

# **Large-Scale DNA Sequencing**

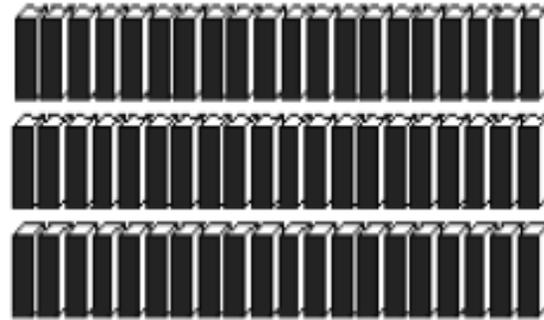
**Current Topics in Genome Analysis, 2000**

**Jeff Touchman, Ph.D.  
Director, Sequence Production Group  
NIH Intramural Sequencing Center  
Tel: 301-435-6156  
Fax: 301-435-6170  
Email: [jefft@nhgri.nih.gov](mailto:jefft@nhgri.nih.gov)**

# Genome Sizes

**Human Genome**

**Mouse Genome**



**~3,000,000,000 bp**

**Fruit Fly Genome**



**~160,000,000 bp**

**Nematode Genome**



**~100,000,000 bp**

**Yeast Genome**



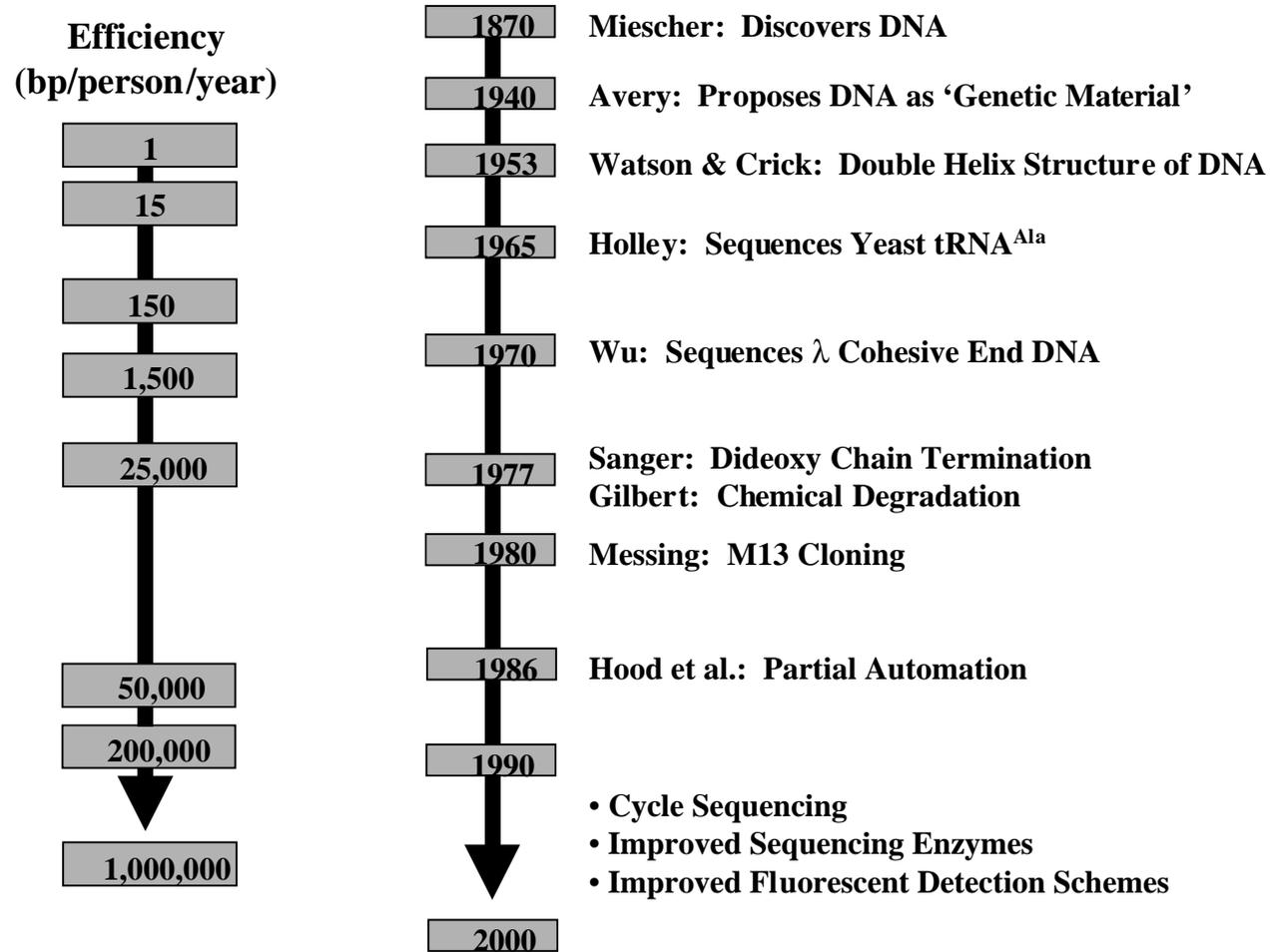
**~15,000,000 bp**

***E. coli* Genome**



**~5,000,000 bp**

# History of DNA Sequencing



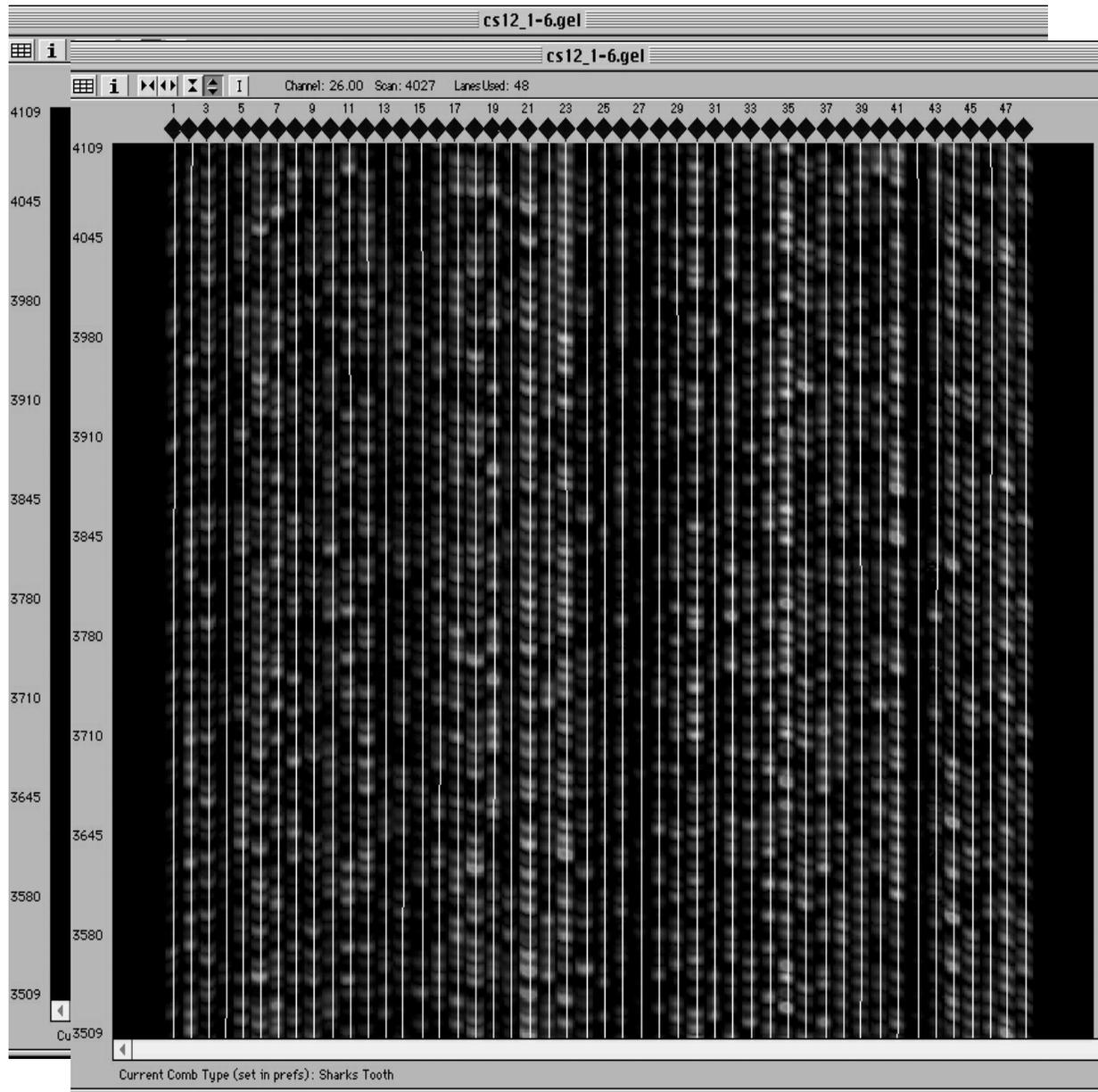
# Radioactive Sequencing



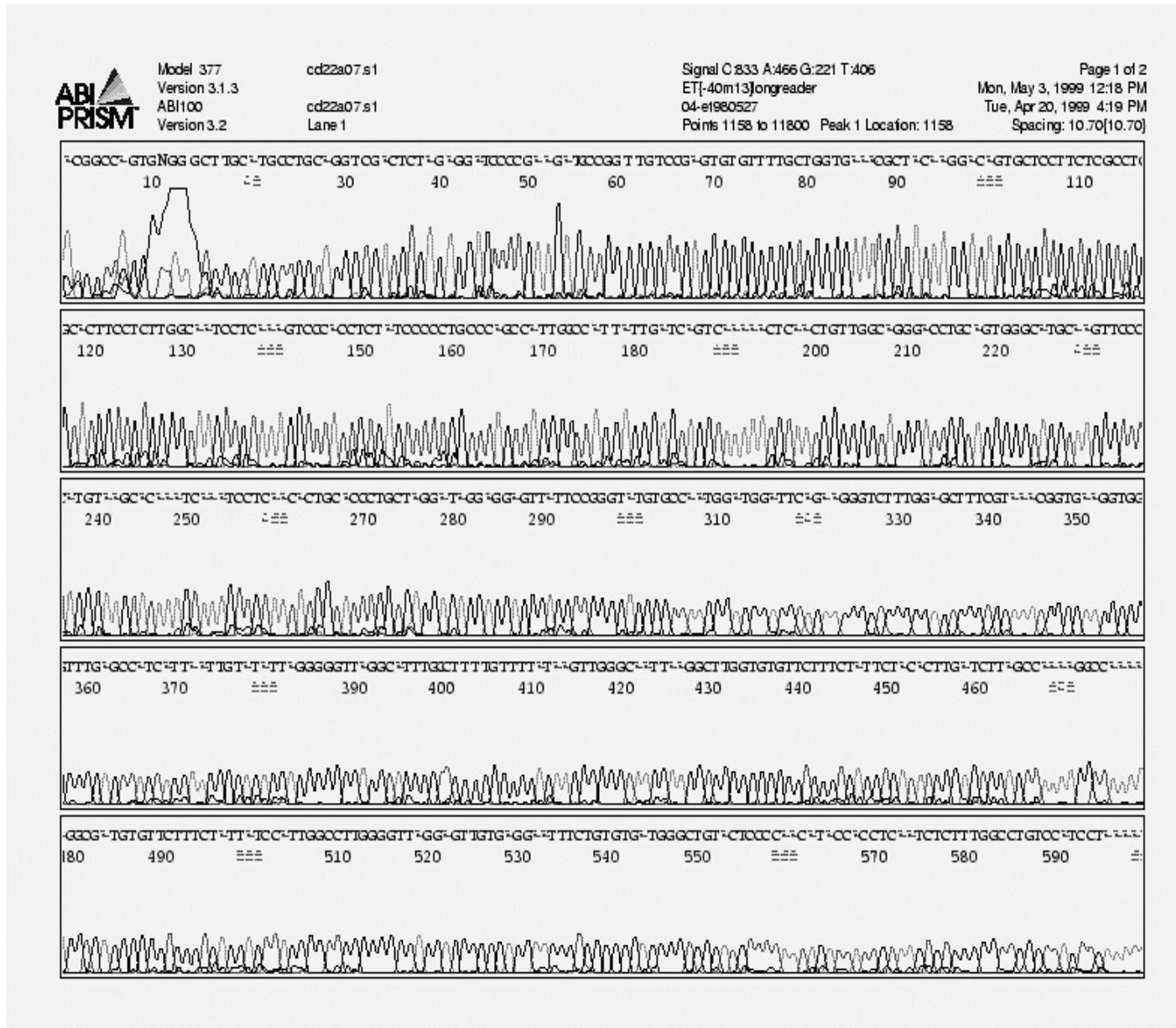
# Perkin Elmer/Applied Biosystems 377



# Fluorescent DNA Sequencing: Lane Tracking



# Fluorescent DNA Sequence Trace



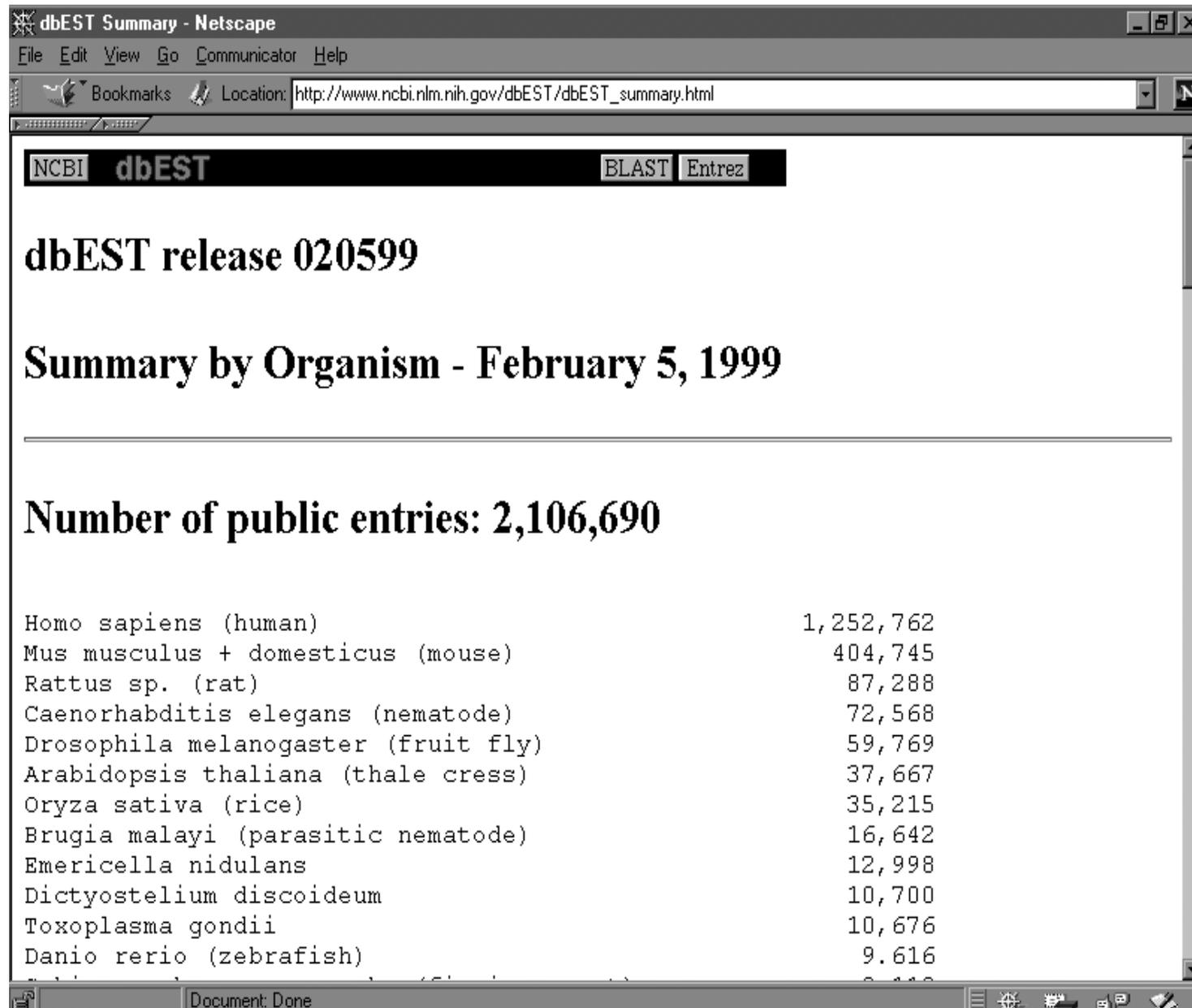
# Expressed-Sequence Tags (ESTs)

- **Single-Pass Sequence of Random cDNA Clone**
- **Often from Normalized cDNA Libraries**



- **3' ESTs More Likely to be Unique Among Gene Family Members**
- **5' ESTs More Likely to Yield Homology Information Indicative of Gene Function**

# Publicly Available ESTs



dbEST Summary - Netscape

File Edit View Go Communicator Help

Bookmarks Location: [http://www.ncbi.nlm.nih.gov/dbEST/dbEST\\_summary.html](http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html)

NCBI dbEST BLAST Entrez

## dbEST release 020599

### Summary by Organism - February 5, 1999

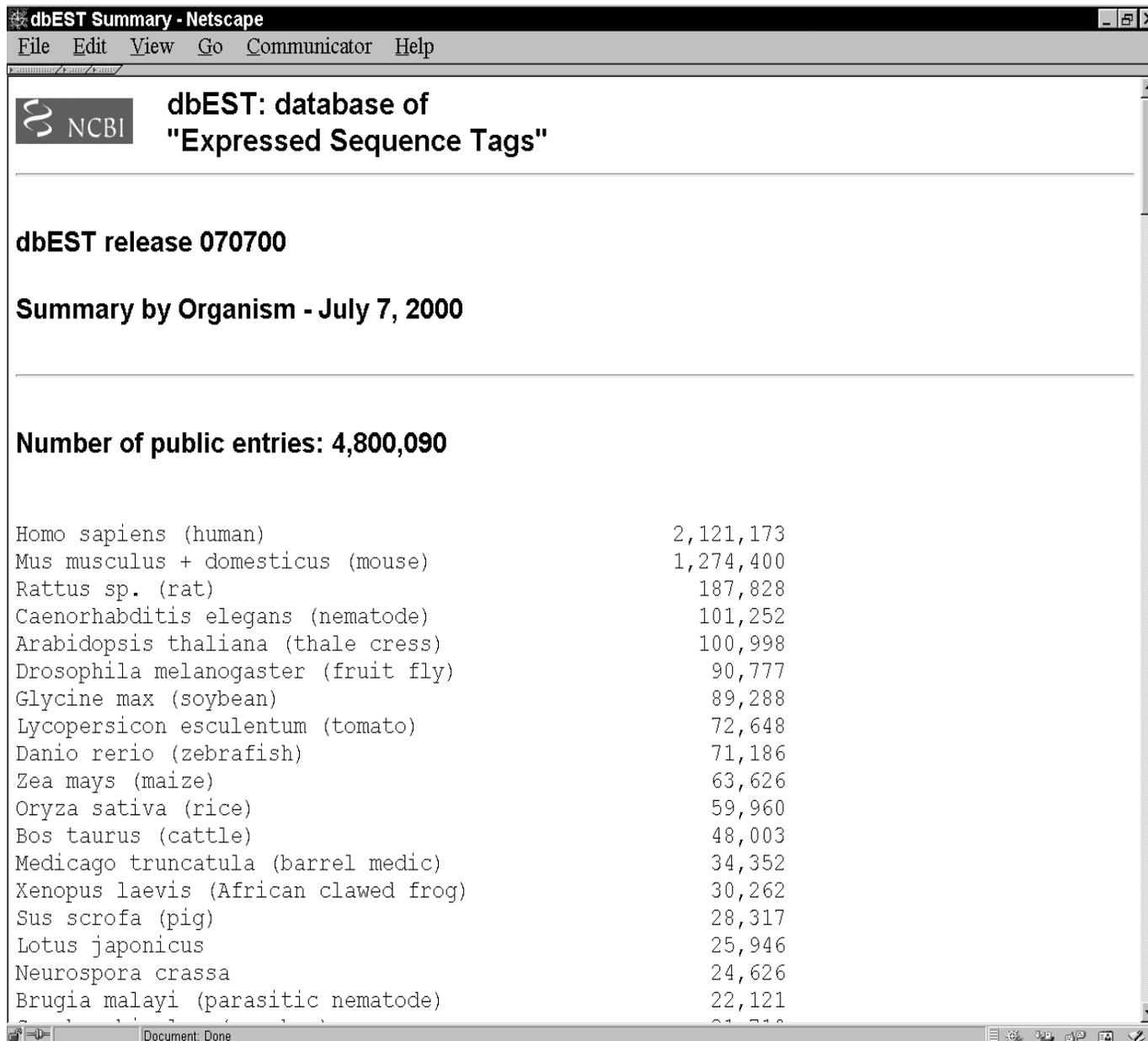
---

**Number of public entries: 2,106,690**

Homo sapiens (human)	1,252,762
Mus musculus + domesticus (mouse)	404,745
Rattus sp. (rat)	87,288
Caenorhabditis elegans (nematode)	72,568
Drosophila melanogaster (fruit fly)	59,769
Arabidopsis thaliana (thale cress)	37,667
Oryza sativa (rice)	35,215
Brugia malayi (parasitic nematode)	16,642
Emericella nidulans	12,998
Dictyostelium discoideum	10,700
Toxoplasma gondii	10,676
Danio rerio (zebrafish)	9,616

Document: Done

# Publicly Available ESTs



dbEST Summary - Netscape

File Edit View Go Communicator Help

 **dbEST: database of  
"Expressed Sequence Tags"**

---

**dbEST release 070700**

**Summary by Organism - July 7, 2000**

---

**Number of public entries: 4,800,090**

Homo sapiens (human)	2,121,173
Mus musculus + domesticus (mouse)	1,274,400
Rattus sp. (rat)	187,828
Caenorhabditis elegans (nematode)	101,252
Arabidopsis thaliana (thale cress)	100,998
Drosophila melanogaster (fruit fly)	90,777
Glycine max (soybean)	89,288
Lycopersicon esculentum (tomato)	72,648
Danio rerio (zebrafish)	71,186
Zea mays (maize)	63,626
Oryza sativa (rice)	59,960
Bos taurus (cattle)	48,003
Medicago truncatula (barrel medic)	34,352
Xenopus laevis (African clawed frog)	30,262
Sus scrofa (pig)	28,317
Lotus japonicus	25,946
Neurospora crassa	24,626
Brugia malayi (parasitic nematode)	22,121

Document: Done

# RH Mapping-Based Gene Map

The screenshot shows a Netscape browser window titled "GeneMap'98 - Netscape". The address bar contains "http://www.ncbi.nlm.nih.gov/genemap/". The main content area features the NCBI logo and the text "A NEW GENE MAP OF THE HUMAN GENOME GeneMap'98 The International RH Mapping Consortium". Below this is a navigation bar with links for Généthon, Sanger, SHGC, WICGR, WTCHG, EBI, and NCBI. A "Chromosomes:" section lists chromosomes 1 through 22 and X. A "Search for:" input field is present. On the left, a sidebar lists navigation options: Background, RH consortium, STS markers, RH mapping, Mapped genes, Gene distribution, Reference intervals, Error analysis, Disease genes, Using this site, Search using text, and Marker view. The main heading is "A New Gene Map of the Human Genome". Below it, the text reads "The International RH Mapping Consortium". A box labeled "The Book of Life" contains the text "The Human Genome Project is entering". Another box states "This web site is the electronic data supplement". The status bar at the bottom shows "Document: Done".

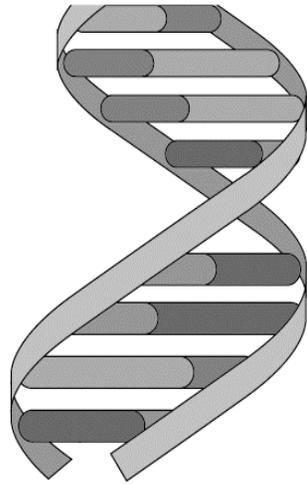
## A Physical Map of 30,000 Human Genes

P. Deloukas,\* G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund, P. Rodriguez-Tomé, L. Hui, T. C. Matisse, K. B. McKusick, J. S. Beckmann, S. Bentolila, M.-T. Bihoreau, B. B. Birren, J. Browne, A. Butler, A. B. Castle, N. Chiannilkulchai, C. Clee, P. J. R. Day, A. Dehejia, T. Dibling, N. Drouot, S. Duprat, C. Fizames, S. Fox, S. Gelling, L. Green, P. Harrison, R. Hocking, E. Holloway, S. Hunt, S. Keil, P. Lijnzaad, C. Louis-Dit-Sully, J. Ma, A. Mendis, J. Miller, J. Morissette, D. Muselet, H. C. Nusbaum, A. Peck, S. Rozen, D. Simon, D. K. Slonim, R. Staples, L. D. Stein, E. A. Stewart, M. A. Suchard, T. Thangarajah, N. Vega-Czarny, C. Webber, X. Wu, J. Hudson, C. Auffray, N. Nomura, J. M. Sikela, M. H. Polymeropoulos, M. R. James, E. S. Lander, T. J. Hudson, R. M. Myers, D. R. Cox, J. Weissenbach, M. S. Boguski, D. R. Bentley

*Science* 282:744-746, 1998

# **The Next Challenge with cDNAs**

- **Construction of Full-Length cDNA Libraries**
- **Identification of Complete Sets of Full-Length cDNA Clones**
- **Sequencing of Complete Sets of Full-Length cDNA Clones**



**M**ammalian  
**G**ene  
**C**ollection

VIEWPOINT

**The Mammalian Gene Collection**

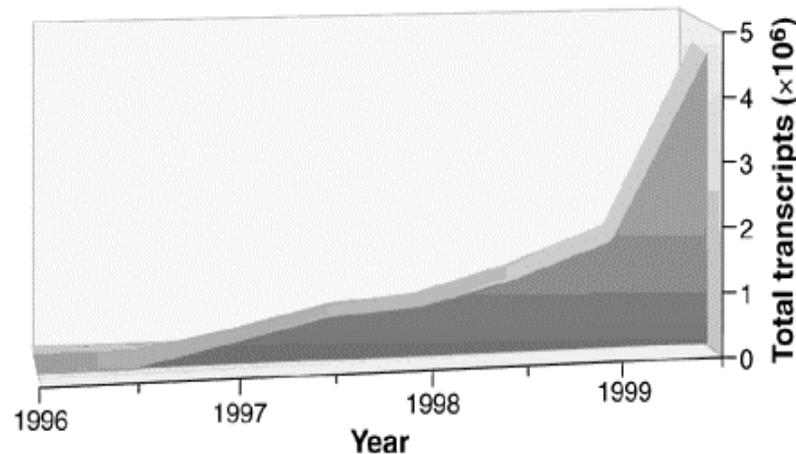
Robert L. Strausberg,<sup>1</sup> Elise A. Feingold,<sup>2</sup> Richard D. Klausner,<sup>1\*</sup> Francis S. Collins,<sup>2\*</sup>

*Science* 286:455-457, 1999

# SAGE

## Serial Analysis of Gene Expression

Designed to take advantage of high-throughput sequencing technology to obtain a quantitative profile of gene expression



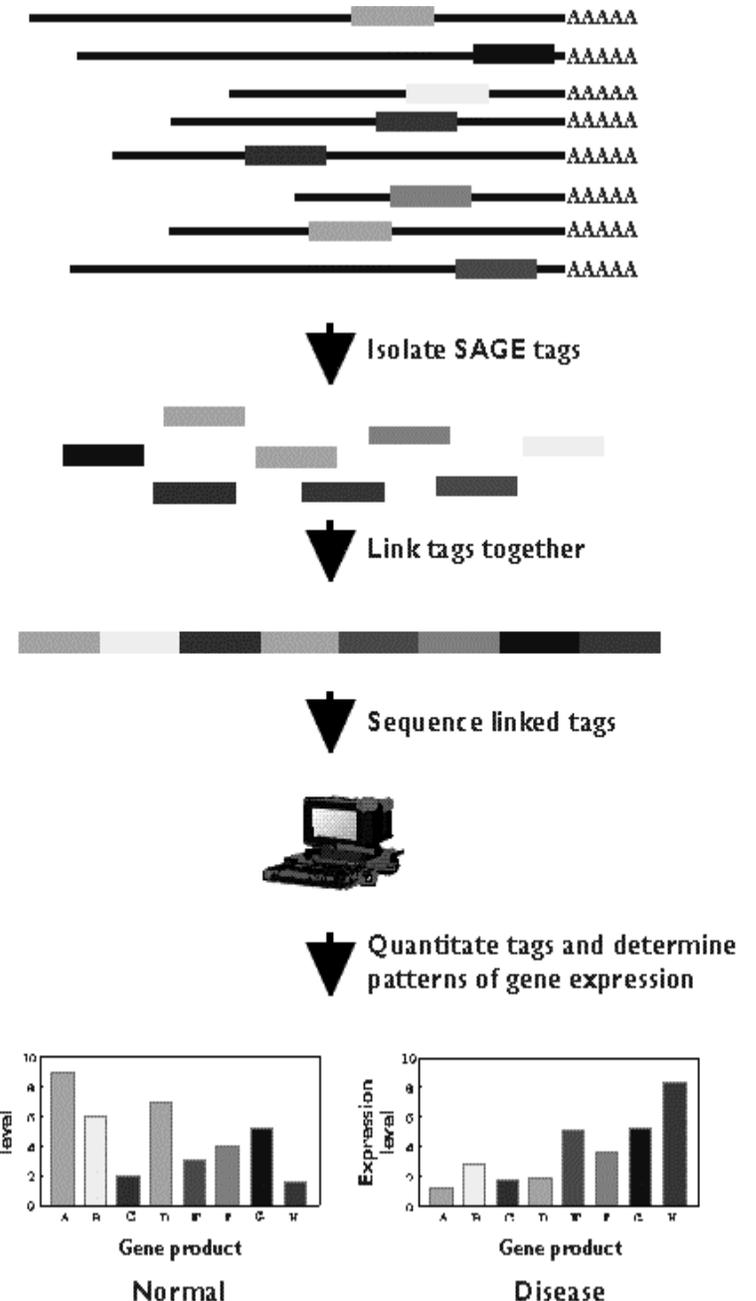
**Velculescu VE et al. *Science* 270: 484-487, 1995**

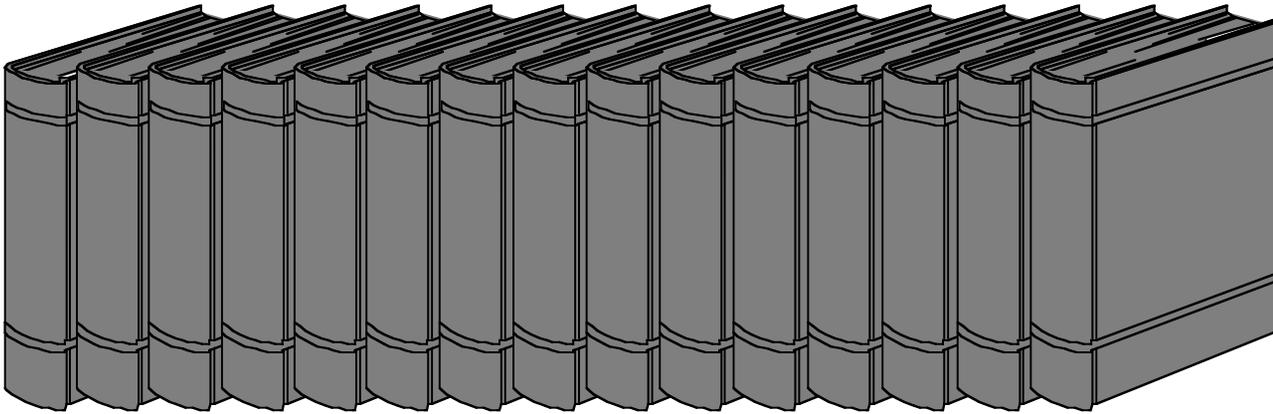
**[www.sagenet.org](http://www.sagenet.org)**

**[www.ncbi.nlm.nih.gov/SAGE/](http://www.ncbi.nlm.nih.gov/SAGE/)**

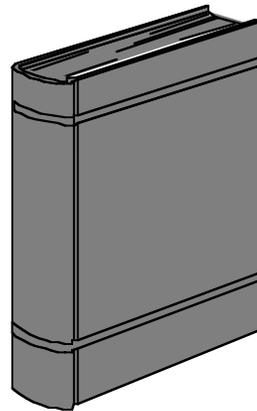
# SAGE Principles

- 1) A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript provided that the tag is obtained from a unique position within each transcript
- 2) Sequence tags can be linked together to form long serial molecules that can be cloned and sequenced
- 3) Quantitation of the number of times a particular tag is observed provides the expression level of the corresponding transcript





**Genome**  
**(~3000 Mb)**



**Chromosome**  
**(~130 Mb)**

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

**BAC**  
**(~0.1-0.2 Mb)**



# Genomic Sequencing: Strategies

- **Transposon-Mediated Sequencing**

**Refined within Drosophila Sequencing Effort**

**Kimmel et al., *Genome Analysis*  
Vol. 1 (CSHL Press)**

- **Shotgun Sequencing**

**Refined within Nematode Sequencing Effort**

**Wilson & Mardis, *Genome Analysis*  
Vol. 1 (CSHL Press)**

# Subclone Construction

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

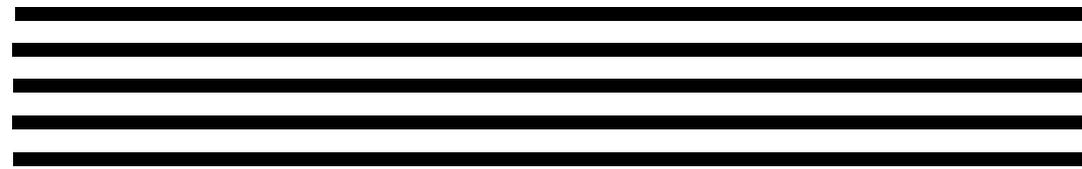
TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

GAGGCCACCCCGCGT
GTGCACGTCCACCACC
```

————— BAC DNA



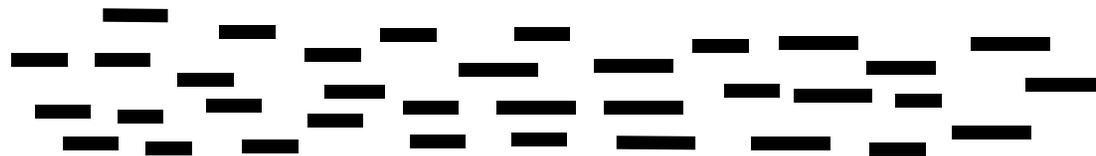
**Prepare Multiple Copies**



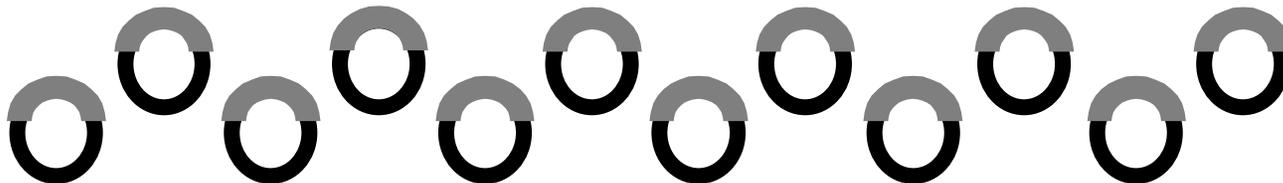
G	G	G	G	GATCGTCTAGAATCTC
G	G	G	G	GAGATCTCTGAGAGTC
G	G	G	G	GTGGGAAACTGTGTGA
T	T	T	T	TGTGACTAGCCACAGT
T	T	T	T	TACGTGTGAGAGATGT
A	A	A	A	ATGATGCACCTGACCC
G	G	G	G	GGGTTTCACTCTCAAC
G	G	G	G	GACTCACTCCACCTCA
G	G	G	G	GAGGCCACCCCGCGT
G	G	G	G	GTGCACGTCCACCACC



**Randomly Fragment**



**Subclone Fragments**



# Poisson calculations

The sequencing strategy for the shotgun approach follows the Lander and Waterman application of the Poisson distribution

The probability a base is not sequenced is given by:

$$P_0 = e^{-c}$$

Where:

- < **c = fold sequence coverage (c=LN/G),**
- < **LN = # bases sequenced, i.e. L = average sequencing read length and N = # reads**
- < **G = target sequence length**
- < **e = 2.718 (e=2.718281828459)**

<b>Fold Coverage</b>	<b><math>P_0 = e^{-c}</math></b>	<b>% not sequenced</b>	<b>% sequenced</b>
<b>1</b>	<b>0.37</b>	<b>37%</b>	<b>63%</b>
<b>2</b>	<b>0.135</b>	<b>13.5%</b>	<b>87.5%</b>
<b>3</b>	<b>0.05</b>	<b>5%</b>	<b>95%</b>
<b>4</b>	<b>0.018</b>	<b>1.8%</b>	<b>98.2%</b>
<b>5</b>	<b>0.0067</b>	<b>0.6%</b>	<b>99.4%</b>
<b>6</b>	<b>0.0025</b>	<b>0.25%</b>	<b>99.75%</b>
<b>7</b>	<b>0.0009</b>	<b>0.09%</b>	<b>99.91%</b>
<b>8</b>	<b>0.0003</b>	<b>0.03%</b>	<b>99.97%</b>
<b>9</b>	<b>0.0001</b>	<b>0.01%</b>	<b>99.99%</b>
<b>10</b>	<b>0.000045</b>	<b>0.005%</b>	<b>99.995%</b>

# Total Gap Length

$$\text{Total Gap Length (bp)} = G e^{-c}$$

Where:

- < c = fold coverage
- < G = target sequence length
- <  $e^{-c} = P_0$

Genome size =	50 kb	150 kb	300 kb	2 Mb	4 Mb
Fold coverage	$G e^{-c}$				
1	18,500	55,500	111,000	740,000	1,480,000
2	6,750	20,250	40,500	270,000	540,000
3	2,500	7,500	15,000	100,000	200,000
4	900	2,700	5,400	36,000	72,000
5	335	1,005	2,010	13,400	26,800
6	125	375	750	5,000	10,000
7	45	135	270	1,800	3,600
8	15	45	90	600	1,200
9	5	15	30	200	400
10	2	6	12	90	180

# Total Number of Gaps

Total number of gaps =  $Ne^{-c}$

Where:

$\langle N = Gc/L =$  number of reads for x-fold coverage

**G** = Target sequence length

**c** = Fold Coverage

**L** = Average sequencing read length

$\langle e^{-c} = P_0$

**50 kb Target Clone:**

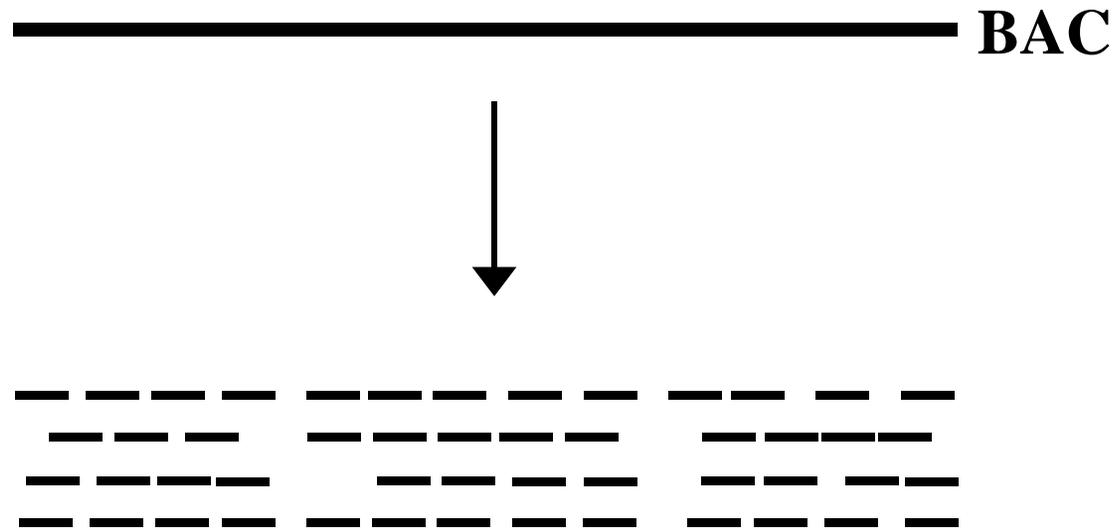
Read length Fold Cov.	400			500			600		
	N	$e^{-c}$	#Gaps = $Ne^{-c}$	N	$e^{-c}$	#Gaps = $Ne^{-c}$	N	$e^{-c}$	#Gaps = $Ne^{-c}$
1	125	0.37	46	100	0.37	37	84	0.37	31
2	250	0.135	34	200	0.135	27	168	0.135	23
3	375	0.05	19	300	0.05	15	242	0.05	12
4	500	0.018	9	400	0.018	7	326	0.018	6
5	625	0.0067	4	500	0.0067	3	410	0.0067	3
6	750	0.0025	2	600	0.0025	2	500	0.0025	1
7	875	0.0009	1	700	0.0009	1	583	0.0009	1
8	1000	0.0003	0	800	0.0003	0	667	0.0003	0
9	1125	0.0001	0	900	0.0001	0	750	0.0001	0
10	1250	0.000045	0	1000	0.000045	0	833	0.000045	0

The values for each fold coverage for a 150kb BAC (G=150,000) with average read length of 500 bases are:

Fold coverage	Total bases sequenced	$e^{-c}$	Total gap length_in bases = $Ge^{-c}$	Number of Gaps = $Ne^{-c}$	Gap Length/# gaps = # bases per gap	% complete
1	150000	0.37	55,500	111	500	63
2	300000	0.135	20,250	81	250	87.5
3	450000	0.05	7,500	45	167	95
4	600000	0.018	2,700	22	123	98.2
5	750000	0.0067	1,005	10	101	99.4
6	900000	0.0025	375	5	75	99.75
7	1050000	0.0009	135	2	68	99.91
8	1200000	0.0003	45	1	45	99.97
9	1350000	0.0001	15	1	15	99.99
10	1500000	0.000045	6	1	6	99.995

For more calculations, see [http://www.genome.ou.edu/poisson\\_calc.html](http://www.genome.ou.edu/poisson_calc.html)

# Shotgun Sequencing Strategy



# **Sequence Assembly Software**

**DNA Star**

**Sequencher (Gene Codes)**

**Assembler (PE/ABI)**

**Gelassemble (GCG)**

**XBAP/XGAP (Staden)**

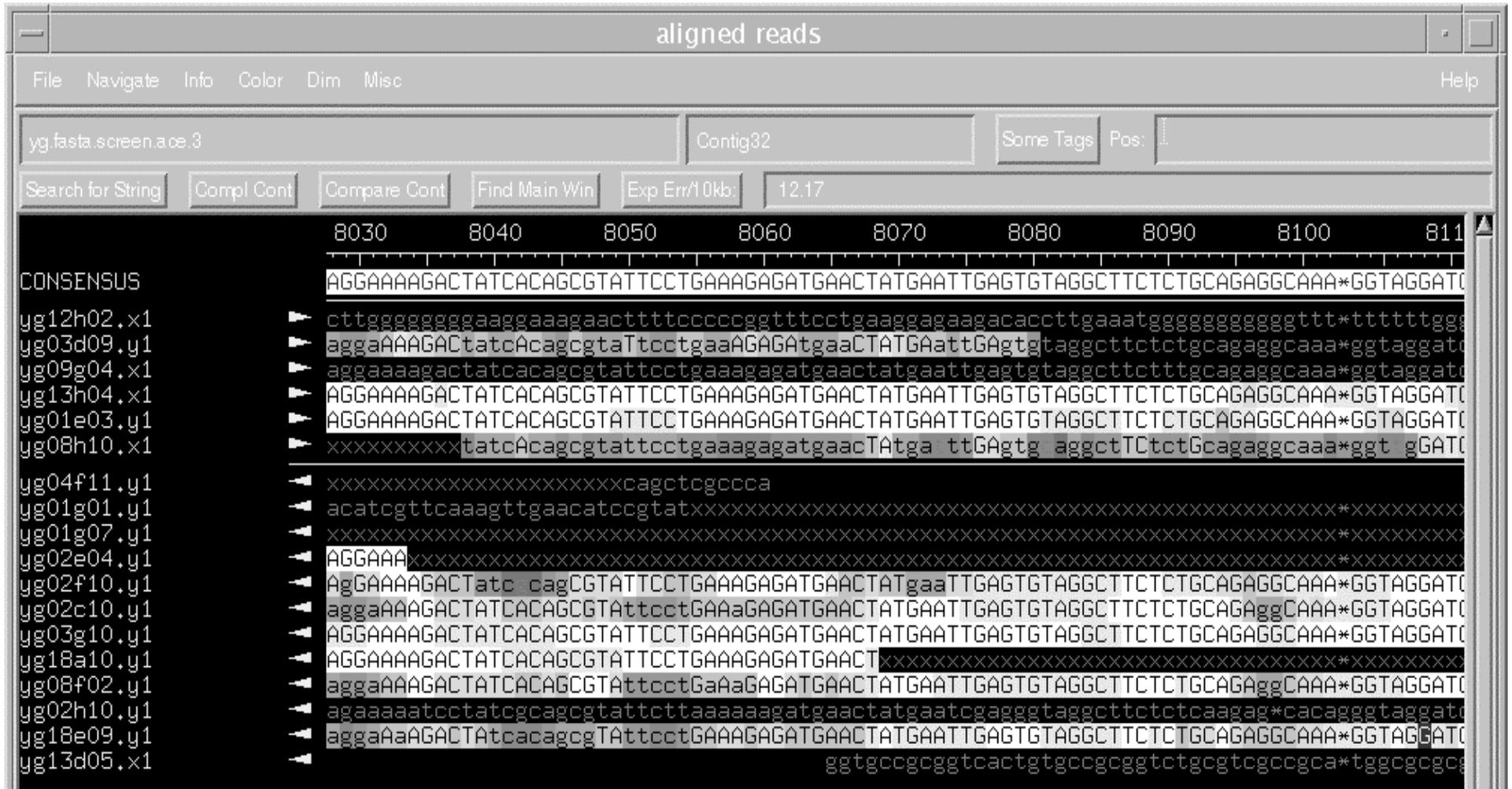
**Phrap (Green)**

# Shotgun Sequence Assembly

The screenshot displays the 'aligned reads' software interface. At the top, the window title is 'aligned reads'. Below the title bar is a menu bar with 'File', 'Navigate', 'Info', 'Color', and 'Help'. A text field contains 'bf.fasta.screen.ace.5' and a 'Contig6' label is visible on the right. Below these are buttons for 'Save Assembly', 'Comp Contig', 'Compare Contigs', and 'Create Exp'. On the right side, there are fields for 'Exp Err/10kb' (value: 3.09) and 'Pos:'. The main display area shows a sequence alignment with a consensus line at the top. The consensus sequence is: CTGCTTCCGAGAGCTTATGATCTCAGAAGCATCTTCCACATCTTGCAAAGCATTATCTACAGGCTGCTATCACTCCCTCGGAGATCA. Below the consensus, several individual reads are listed with their identifiers and alignment status (indicated by arrows):  
bf04h01.y1 CTGCTTCCGAGAGCTTATGATCTCAGaaGCATCTTCCACATCTTGCAAAGCATTATCTACAggCTgctATCACTccCTCggagatca  
bf01b06.f1 CTGCTTCCGAGAGCTTATGATCTCAGAAGCATcttcCACATCTTGCAAAGCATTATCTACAGGCTGCTATCACTCcctcgGAGATca  
bf19c10.y1 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxCAGAAGCATCTTCCACATCTTGCAAAGCATTATCTACAGGCTGCTATCACTCCCTCGGAGATCA  
bf20b11.x1 xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxTTATGATCTCAGAAGCATCTTCCACATCTTGCAAAGCATTATCTACAGGCTGCTATCACTCCCTCGGAGATCA  
bf22g12.y1 cTGCTtccGAGAGCTTATGATCTCaGAAGCATCTTCCACATCTTGCAAAGCATTATCTcACAGGCTGCTATCACTCCCTCGGAGATCA  
bf02h06.f1 CTGCTTCCGAGAGCTTATGATCTCAGAAGCATCTTCCACATCTTGCAAAGCATTATCTACAGGCTGCTATCACTCCCTCGGAGATCA  
bf22e07.f1 CTGCTTCCgagagcTTATGATCTCagaagcaTCTTCCACATCTTGCAAAGCATTATCTACAGGCTGCTATCACTCCCTCGGAGATCA  
At the bottom of the window, there are navigation arrows and a 'dismiss' button.

“Consed” (Gordon et al., *Genome Research* 8:195-202, 1998)

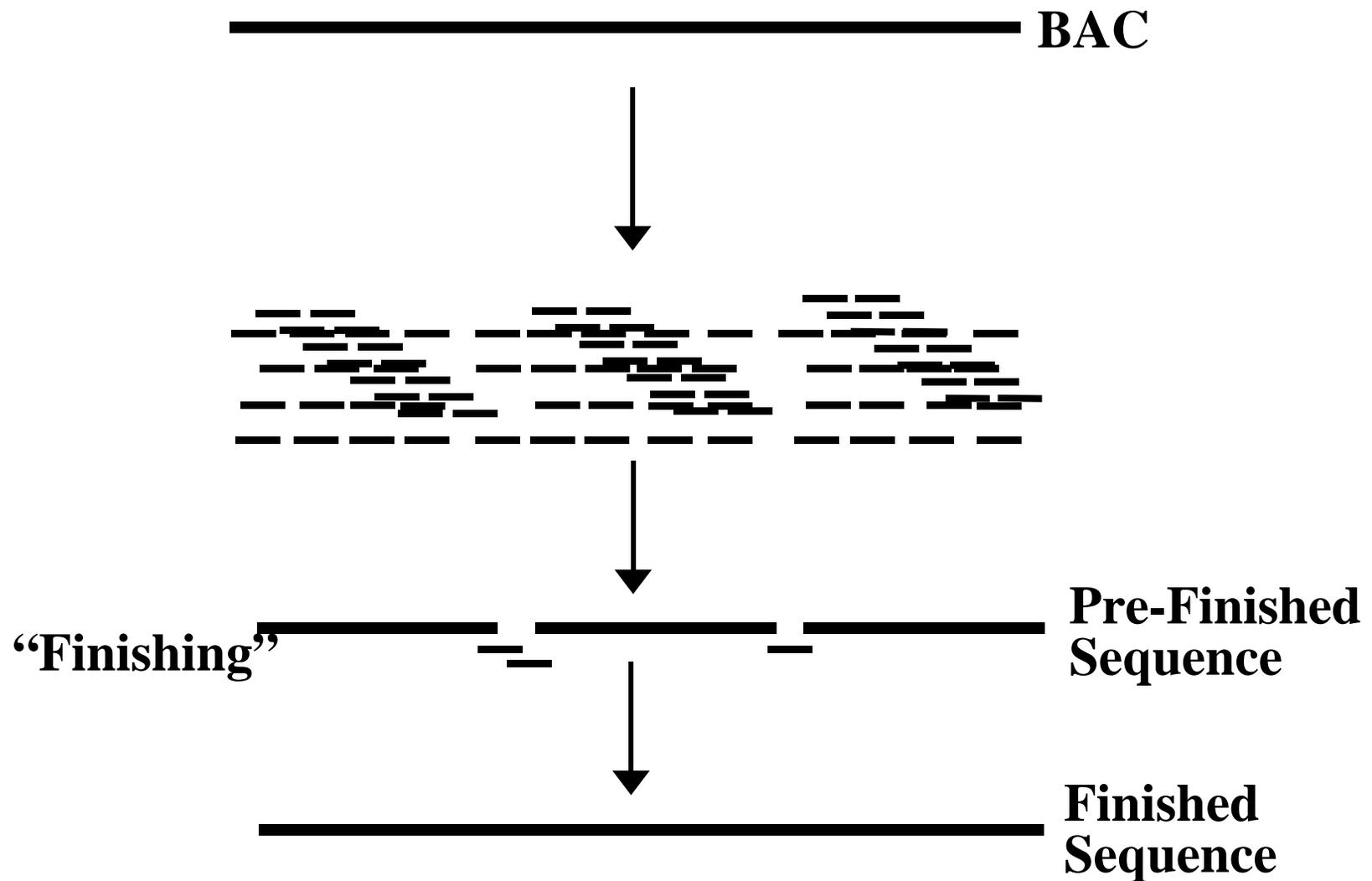
# Shotgun Sequence Assembly



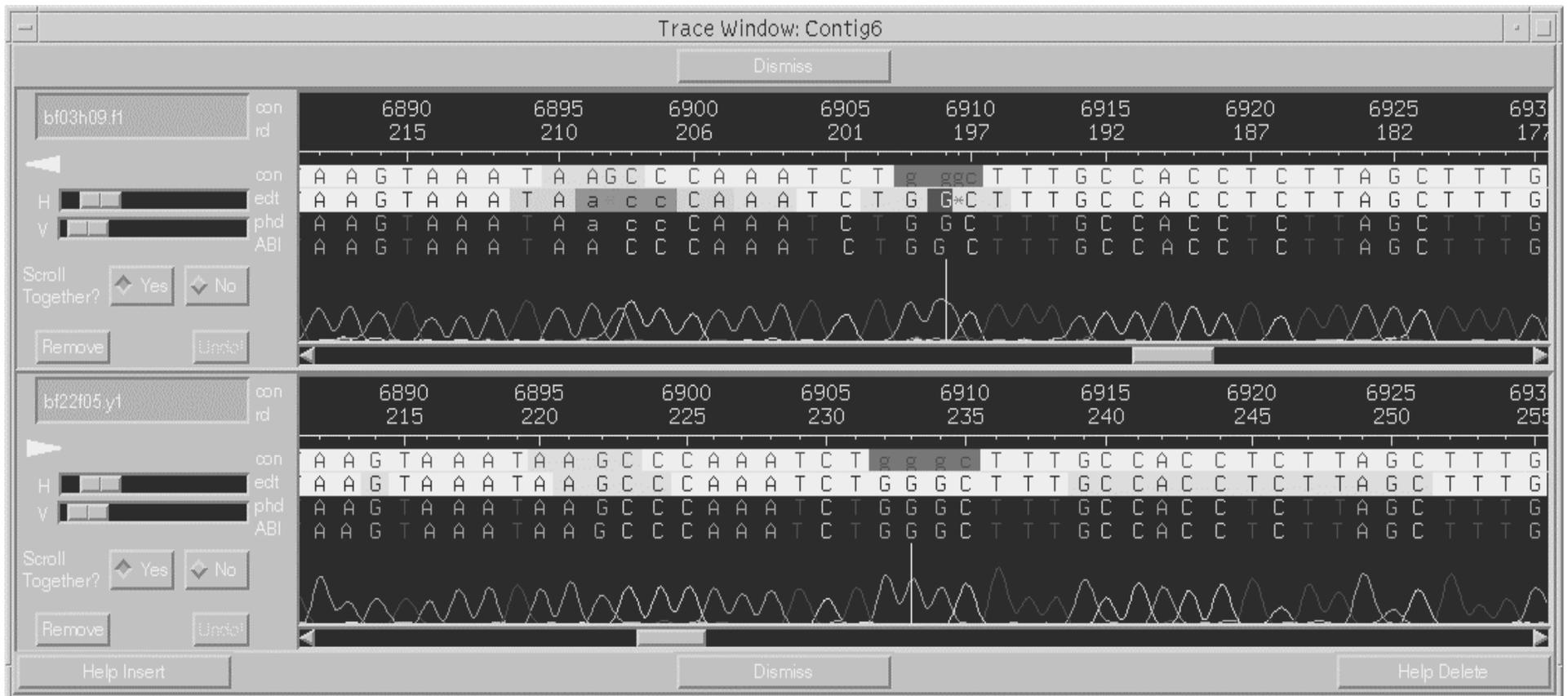
“Consed” (Gordon et al., *Genome Research* 8:195-202, 1998)



# Shotgun Sequencing Strategy



# Resolve Ambiguities...



# **DNA Sequencing in the Human Genome Project**

# Complete Sequences of Microbial Genomes

**TIGR Microbial Database - Netscape**

File Edit View Go Communicator Help



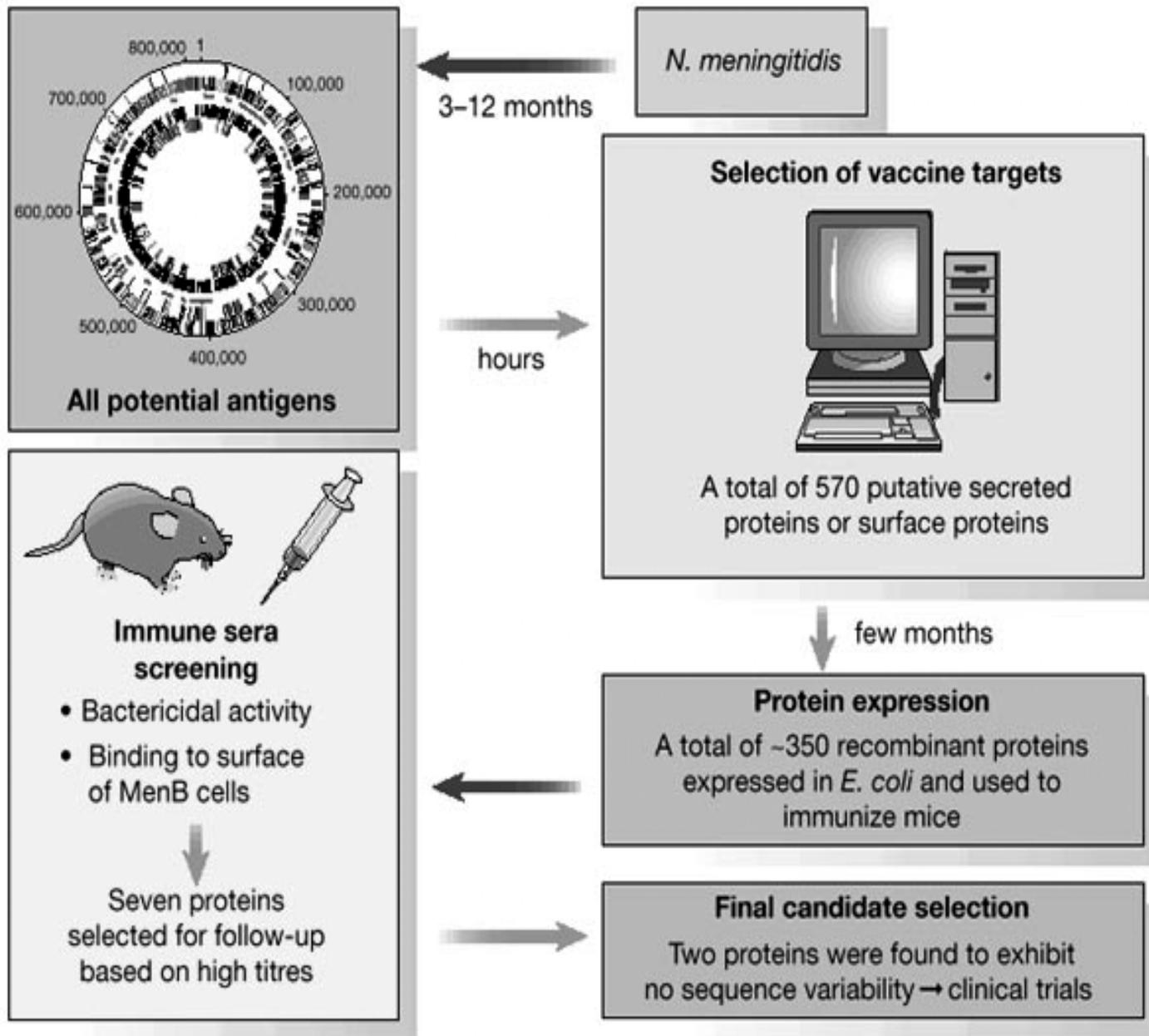
**TIGR Microbial Database:**  
a listing of microbial genomes and chromosomes completed and in progress

Published microbial genomes and chromosomes (scroll down for genomes in progress)

	Link	Genome	Strain	Domain	Size (Mb)	Institution	Funding	Publication
1		<i>Haemophilus influenzae</i> Rd	KW20	B	1.83	TIGR	TIGR	Fleischmann <i>et. al.</i> , <i>Science</i> 269:496-512 (1995)
2		<i>Mycoplasma genitalium</i>	G-37	B	0.58	TIGR	DOE	Fraser <i>et. al.</i> , <i>Science</i> 270:397-403 (1995)
3		<i>Methanococcus jannaschii</i>	DSM 2661	A	1.66	TIGR	DOE	Bult <i>et. al.</i> , <i>Science</i> 273:1058-1073 (1996)
4		<i>Synechocystis</i> sp.	PCC 6803	B	3.57	Kazusa DNA Research Inst.		Kaneko <i>et. al.</i> , <i>DNA Res.</i> 3: 109-136 (1996)
5		<i>Mycoplasma pneumoniae</i>	M129	B	0.81	Univ. of Heidelberg	DFG	Himmelreich <i>et. al.</i> , <i>Nuc. Acid Res.</i> 24:4420-4449 (1996)
6		<i>Saccharomyces cerevisiae</i>	S288C	E	13	International Consortium	EC, NHGRI, Wellcome Trust, McGill U., RIKEN	Goffeau <i>et. al.</i> , <i>Nature</i> 387 (Suppl.) 5-105 (1997)
7		<i>Helicobacter pylori</i>	26695	B	1.66	TIGR	TIGR	Tomb <i>et. al.</i> , <i>Nature</i> 388:539-547 (1997)
8		<i>Escherichia coli</i>	K-12	B	4.60	University of Wisconsin	NHGRI	Blattner <i>et. al.</i> , <i>Science</i> 277:1453-1474 (1997)
9		<i>Methanobacterium thermoautotrophicum</i>	delta H	A	1.75	Genome Therapeutics & Ohio State Univ.	DOE	Smith <i>et. al.</i> , <i>J. Bacteriology</i> , 179:7135-7155 (1997)
10		<i>Bacillus subtilis</i>	168	B	4.20	International Consortium	EC	Kunst <i>et. al.</i> , <i>Nature</i> 390: 249-256 (1997)

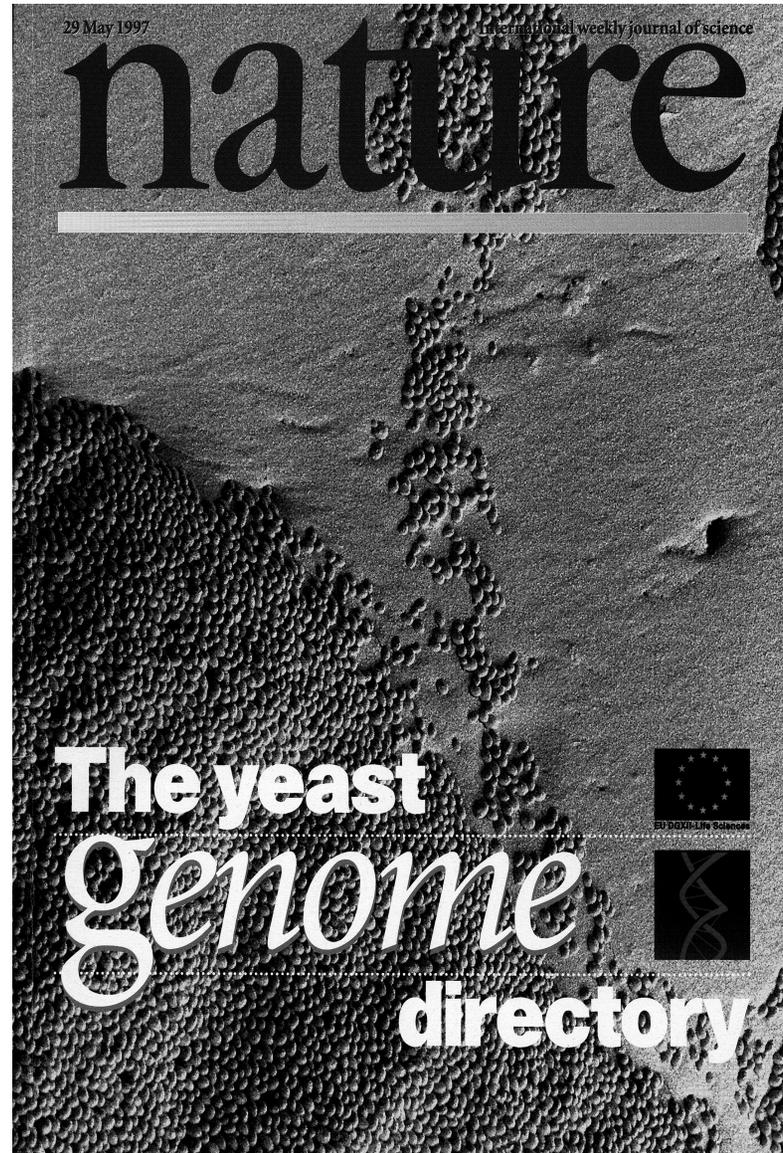
Document: Done

<http://www.tigr.org/tdb/mdb>



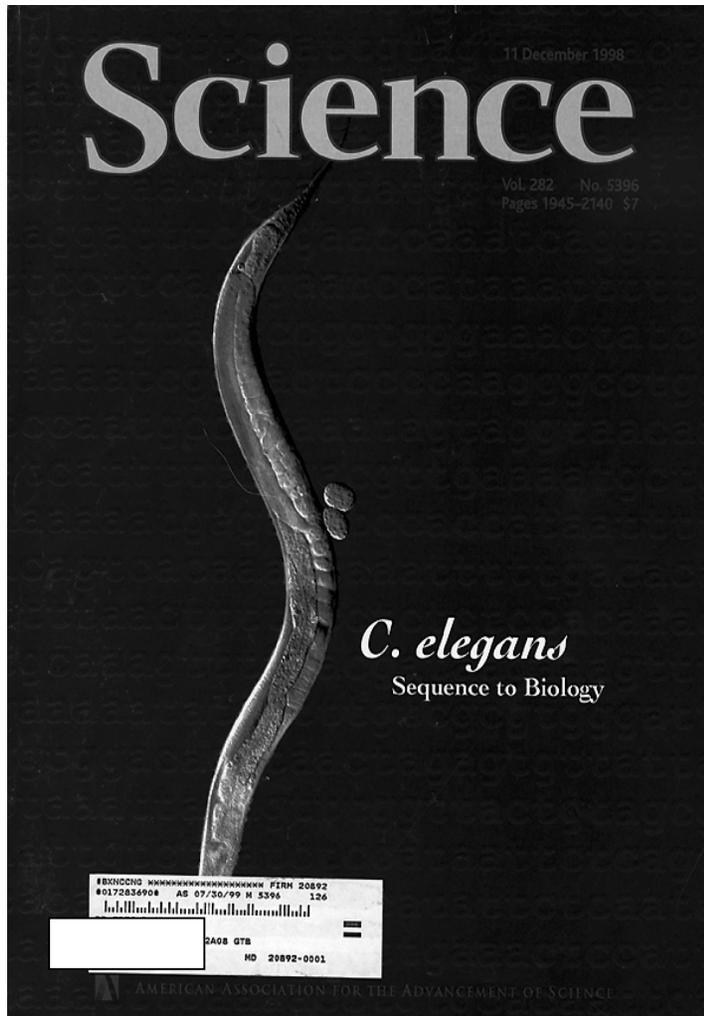
Fraser et al., *Nature* 407: 799-803,

# First Eukaryotic Genome Sequence



*Nature* 387:1-105, 1997

# First Complete Sequence of Multicellular Organism

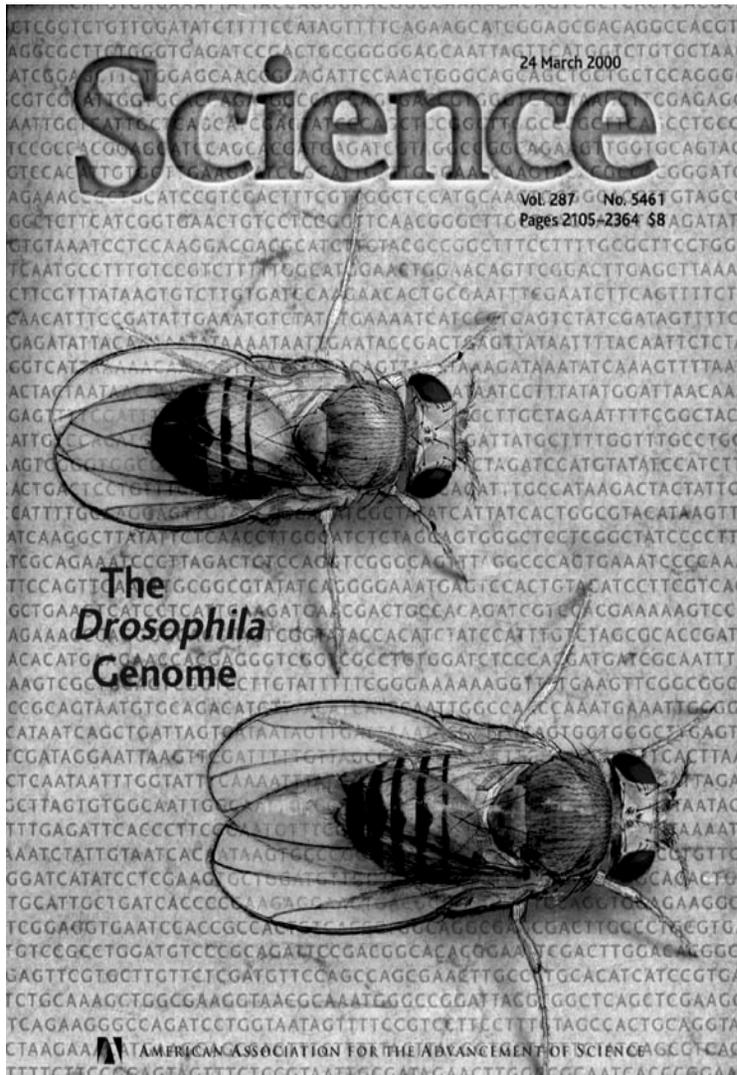


**Genome Sequence of the Nematode *C. elegans*:  
A Platform for Investigating Biology**

The *C. elegans* Sequencing Consortium\*

***Science* 282:1012-2018, 1998**

# Second Animal Genome Sequence



THE DROSOPHILA GENOME  
REVIEW

## The Genome Sequence of *Drosophila melanogaster*

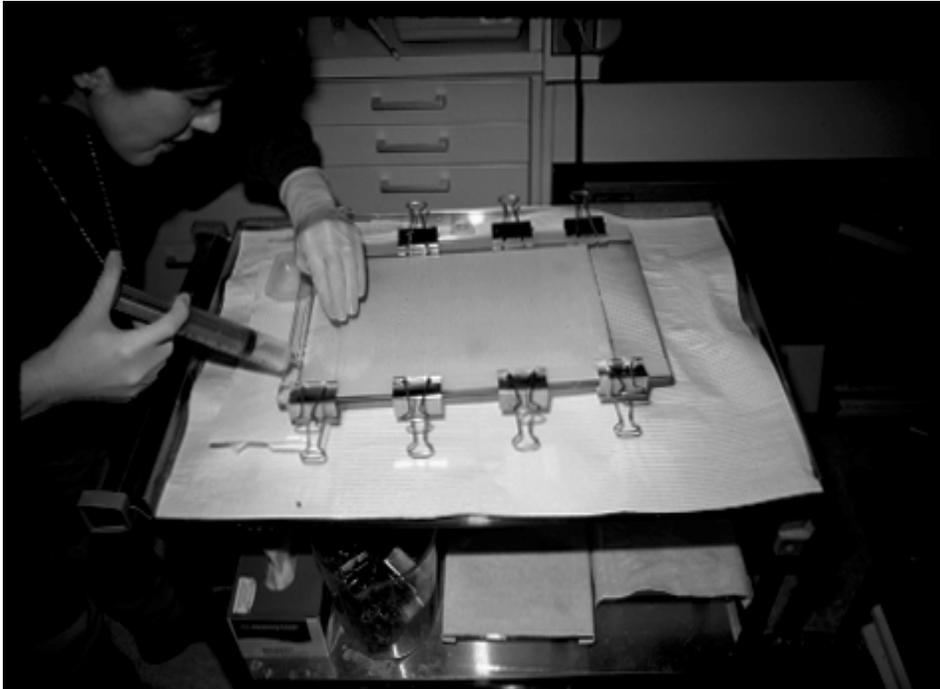
Mark D. Adams,<sup>1\*</sup> Susan E. Celniker,<sup>2</sup> Robert A. Holt,<sup>1</sup> Cheryl A. Evans,<sup>1</sup> Jeannine D. Gocayne,<sup>1</sup> Peter G. Amanatides,<sup>1</sup> Steven E. Scherer,<sup>2</sup> Peter W. Li,<sup>1</sup> Roger A. Hoskins,<sup>2</sup> Richard F. Galle,<sup>2</sup> Reed A. George,<sup>2</sup> Suzanna E. Lewis,<sup>4</sup> Stephen Richards,<sup>2</sup> Michael Ashburner,<sup>3</sup> Scott N. Henderson,<sup>1</sup> Granger G. Sutton,<sup>1</sup> Jennifer R. Wortman,<sup>1</sup> Mark D. Yandell,<sup>1</sup> Qing Zhang,<sup>1</sup> Lin X. Chen,<sup>1</sup> Rhonda C. Brandon,<sup>1</sup> Yu-Hui C. Rogers,<sup>1</sup> Robert G. Blazej,<sup>2</sup> Mark Champe,<sup>2</sup> Barret D. Pfeiffer,<sup>2</sup> Kenneth H. Wan,<sup>2</sup> Clare Doyle,<sup>2</sup> Evan G. Baxter,<sup>2</sup> Gregg Helt,<sup>4</sup> Catherine R. Nelson,<sup>4</sup> George L. Gabor Miklos,<sup>7</sup> Josep F. Abril,<sup>8</sup> Anna Agbayani,<sup>2</sup> Hui-Jin An,<sup>1</sup> Cynthia Andrews-Pfannkoch,<sup>1</sup> Danita Baldwin,<sup>1</sup> Richard M. Ballew,<sup>1</sup> Anand Basu,<sup>1</sup> James Baxendale,<sup>1</sup> Leyla Bayraktaroglu,<sup>9</sup> Ellen M. Beasley,<sup>1</sup> Karen Y. Beeson,<sup>1</sup> P. V. Benos,<sup>10</sup> Benjamin P. Berman,<sup>2</sup> Deepali Bhandari,<sup>1</sup> Slava Bolshakov,<sup>11</sup> Dana Borkova,<sup>12</sup> Michael R. Botchan,<sup>13</sup> John Bouck,<sup>3</sup> Peter Brokstein,<sup>4</sup> Phillippe Brottier,<sup>14</sup> Kenneth C. Burtis,<sup>15</sup> Dana A. Busam,<sup>1</sup> Heather Butler,<sup>16</sup> Edouard Cadieu,<sup>17</sup> Angela Center,<sup>17</sup> Ishwar Chandra,<sup>1</sup> J. Michael Cherry,<sup>18</sup> Simon Cawley,<sup>19</sup> Carl Dahlke,<sup>1</sup> Lionel B. Davenport,<sup>1</sup> Peter Davies,<sup>1</sup> Beatriz de Pablos,<sup>20</sup> Arthur Delcher,<sup>1</sup> Zuoming Deng,<sup>1</sup> Anne Deslattes Mays,<sup>1</sup> Ian Dew,<sup>1</sup> Suzanne M. Dietz,<sup>1</sup> Kristina Dodson,<sup>1</sup> Lisa E. Doup,<sup>1</sup> Michael Downes,<sup>21</sup> Shannon Dugan-Rocha,<sup>3</sup> Boris C. Dunkov,<sup>22</sup> Patrick Dunn,<sup>1</sup> Kenneth J. Durbin,<sup>3</sup> Carlos C. Evangelista,<sup>1</sup> Concepcion Ferraz,<sup>23</sup> Steven Ferriera,<sup>1</sup> Wolfgang Fleischmann,<sup>3</sup> Carl Fosler,<sup>1</sup> Andrei E. Gabrielian,<sup>1</sup> Neha S. Garg,<sup>1</sup> William M. Gelbart,<sup>9</sup> Ken Glasser,<sup>1</sup> Anna Glodek,<sup>1</sup> Fangcheng Gong,<sup>1</sup> J. Harley Gorrell,<sup>3</sup> Zhiping Gu,<sup>1</sup> Ping Guan,<sup>1</sup> Michael Harris,<sup>1</sup> Nomi L. Harris,<sup>2</sup> Damon Harvey,<sup>4</sup> Thomas J. Heiman,<sup>1</sup> Judith R. Hernandez,<sup>3</sup> Jarrett Houck,<sup>1</sup> Damon Hostin,<sup>1</sup> Kathryn A. Houston,<sup>2</sup> Timothy J. Howland,<sup>1</sup> Ming-Hui Wei,<sup>1</sup> Chinyere Ibegwam,<sup>1</sup> Mena Jalali,<sup>1</sup> Francis Kalush,<sup>1</sup> Gary H. Karpen,<sup>21</sup> Zhaoxi Ke,<sup>1</sup> James A. Kennison,<sup>24</sup> Karen A. Ketchum,<sup>1</sup> Bruce E. Kimmel,<sup>2</sup> Chinnappa D. Kodira,<sup>1</sup> Cheryl Kraft,<sup>1</sup> Saul Kravitz,<sup>1</sup> David Kulp,<sup>6</sup> Zhongwu Lai,<sup>1</sup> Paul Lasko,<sup>25</sup> Yiding Lei,<sup>1</sup> Alexander A. Levitsky,<sup>1</sup> Jiayin Li,<sup>1</sup> Zhenya Li,<sup>1</sup> Yong Liang,<sup>1</sup> Xiaoying Lin,<sup>26</sup> Xiangjun Liu,<sup>1</sup> Bettina Mattel,<sup>1</sup> Tina C. McIntosh,<sup>1</sup> Michael P. McLeod,<sup>3</sup> Duncan McPherson,<sup>1</sup> Gennady Merkulov,<sup>1</sup> Natalia V. Milshina,<sup>1</sup> Clark Mobarry,<sup>1</sup> Joe Morris,<sup>6</sup> Ali Moshrefi,<sup>2</sup> Stephen M. Mount,<sup>27</sup> Mee Moy,<sup>1</sup> Brian Murphy,<sup>1</sup> Lee Murphy,<sup>28</sup> Donna M. Muzny,<sup>3</sup> David L. Nelson,<sup>3</sup> David R. Nelson,<sup>29</sup> Keith A. Nelson,<sup>1</sup> Katherine Nixon,<sup>2</sup> Deborah R. Nusskern,<sup>1</sup> Joanne M. Pacleb,<sup>2</sup> Michael Palazzolo,<sup>2</sup> Gjang S. Pittman,<sup>1</sup> Sue Pan,<sup>1</sup> John Pollard,<sup>1</sup> Vinita Puri,<sup>1</sup> Martin G. Reese,<sup>4</sup> Knut Reinert,<sup>1</sup> Karin Remington,<sup>1</sup> Robert D. C. Saunders,<sup>30</sup> Frederick Scheeler,<sup>1</sup> Hua Shen,<sup>3</sup> Bixiang Christopher Shue,<sup>1</sup> Inga Sidén-Kiamos,<sup>11</sup> Michael Simpson,<sup>1</sup> Marian P. Skupski,<sup>1</sup> Tom Smith,<sup>1</sup> Eugene Spier,<sup>1</sup> Allan C. Spradling,<sup>31</sup> Mark Stapleton,<sup>2</sup> Renee Strong,<sup>1</sup> Eric Sun,<sup>1</sup> Robert Svirskas,<sup>32</sup> Cyndee Tector,<sup>1</sup> Russell Turner,<sup>1</sup> Eli Venter,<sup>1</sup> Aihui H. Wang,<sup>1</sup> Xin Wang,<sup>1</sup> Zhen-Yuan Wang,<sup>1</sup> David A. Wassarman,<sup>23</sup> George M. Weinstock,<sup>3</sup> Jean Weissenbach,<sup>14</sup> Sherita M. Williams,<sup>1</sup> Trevor Woodage,<sup>1</sup> Kim C. Worley,<sup>3</sup> David Wu,<sup>1</sup> Song Yang,<sup>2</sup> Q. Allison Yao,<sup>1</sup> Jane Ye,<sup>1</sup> Ru-Fang Yeh,<sup>10</sup> Jayshree S. Zaveri,<sup>1</sup> Ming Zhan,<sup>1</sup> Guangren Zhang,<sup>1</sup> Qi Zhao,<sup>1</sup> Liansheng Zheng,<sup>1</sup> Xiangqun H. Zheng,<sup>1</sup> Fei N. Zhong,<sup>1</sup> Wenyan Zhong,<sup>1</sup> Xiaojun Zhou,<sup>1</sup> Shiaoqing Zhu,<sup>1</sup> Xiaohong Zhu,<sup>1</sup> Hamilton O. Smith,<sup>1</sup> Richard A. Gibbs,<sup>3</sup> Eugene W. Myers,<sup>1</sup> Gerald M. Rubin,<sup>34</sup> J. Craig Venter<sup>1</sup>

*Science* 287:2185-2195, 2000

# **Limitations of Gel-Based DNA Sequencing**

# Limitations of Gel-Based Systems

## Gel Pouring



## Gel Loading

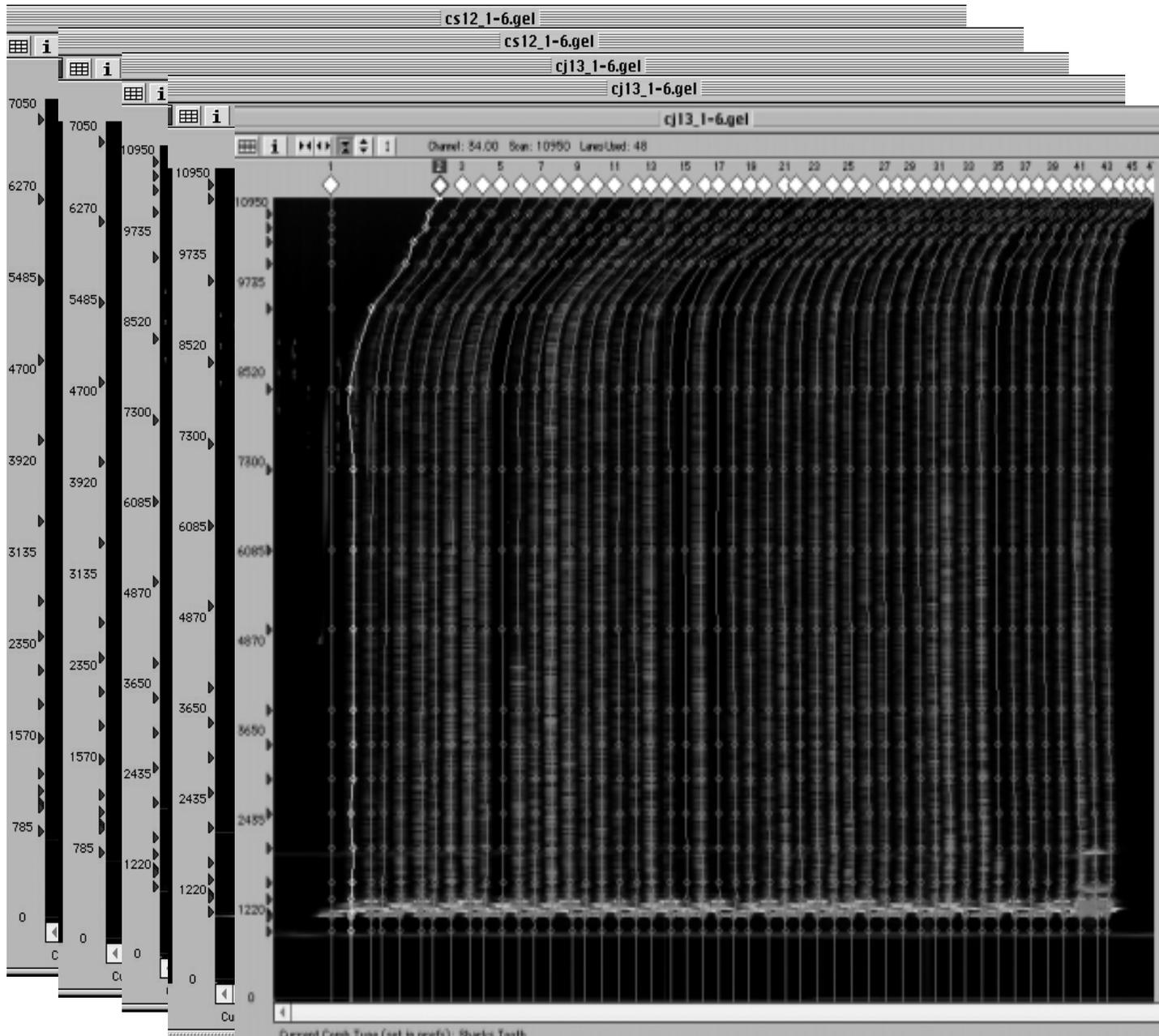


# Lots of Sequence Reads...

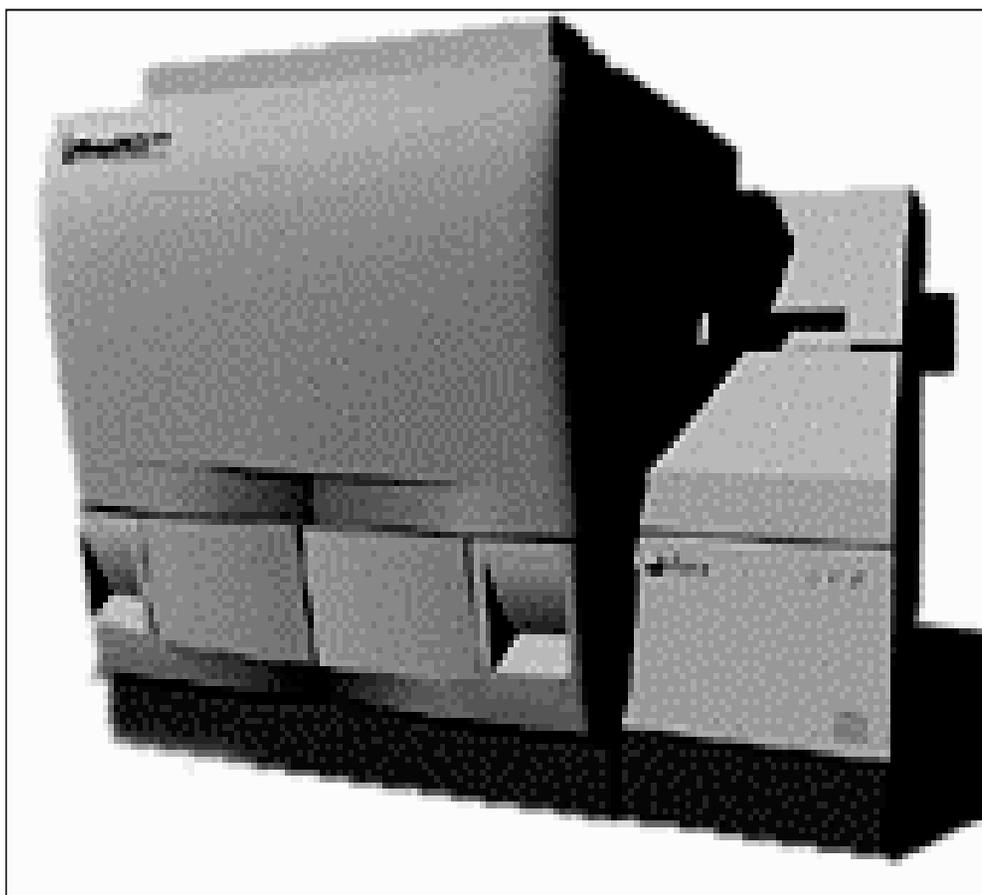


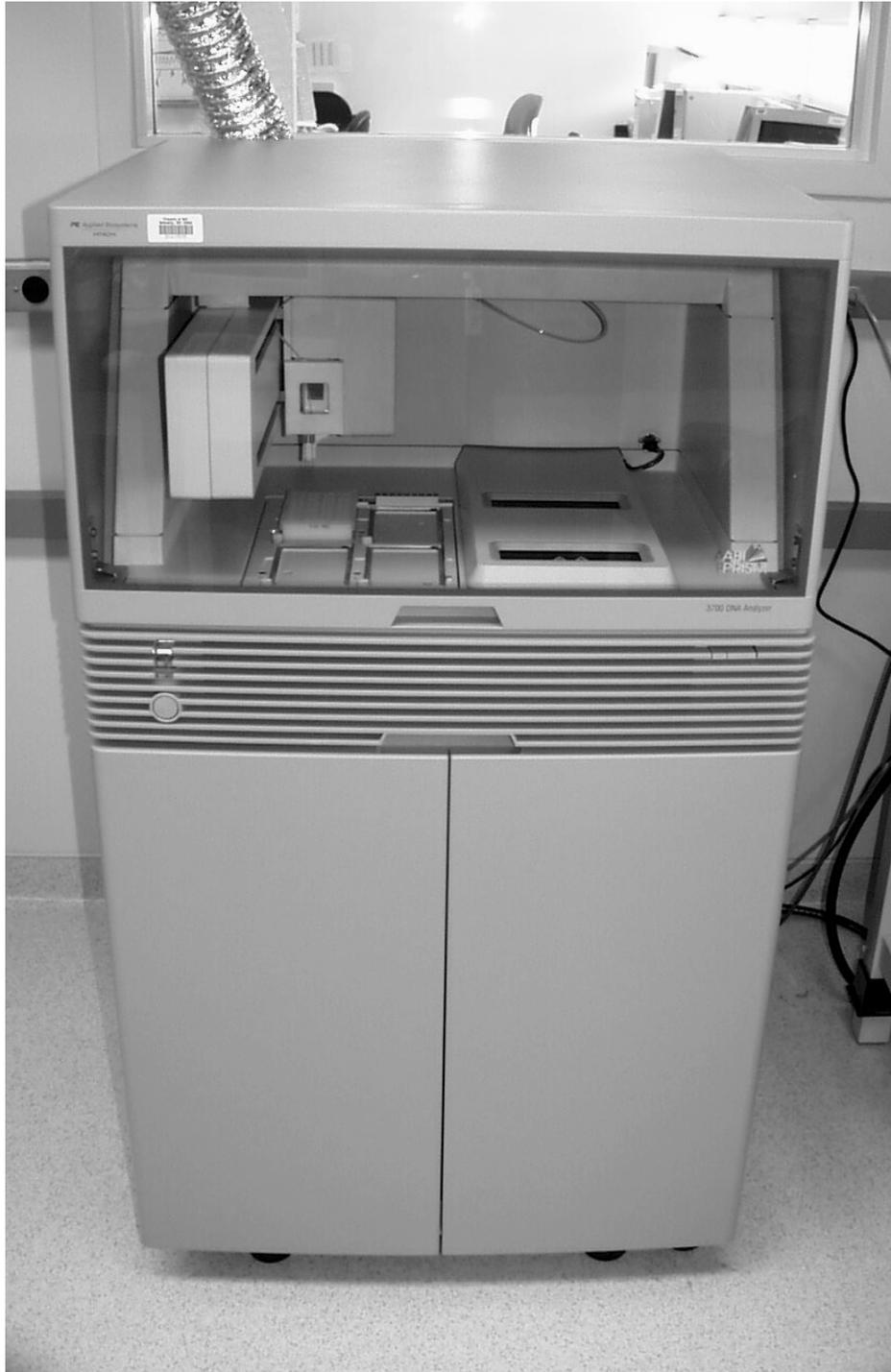
# Lots of Gels!

# Effort of Re-Tracking Gels



# Molecular Dynamics MegaBACE 1000





# Applied Biosystems 3700





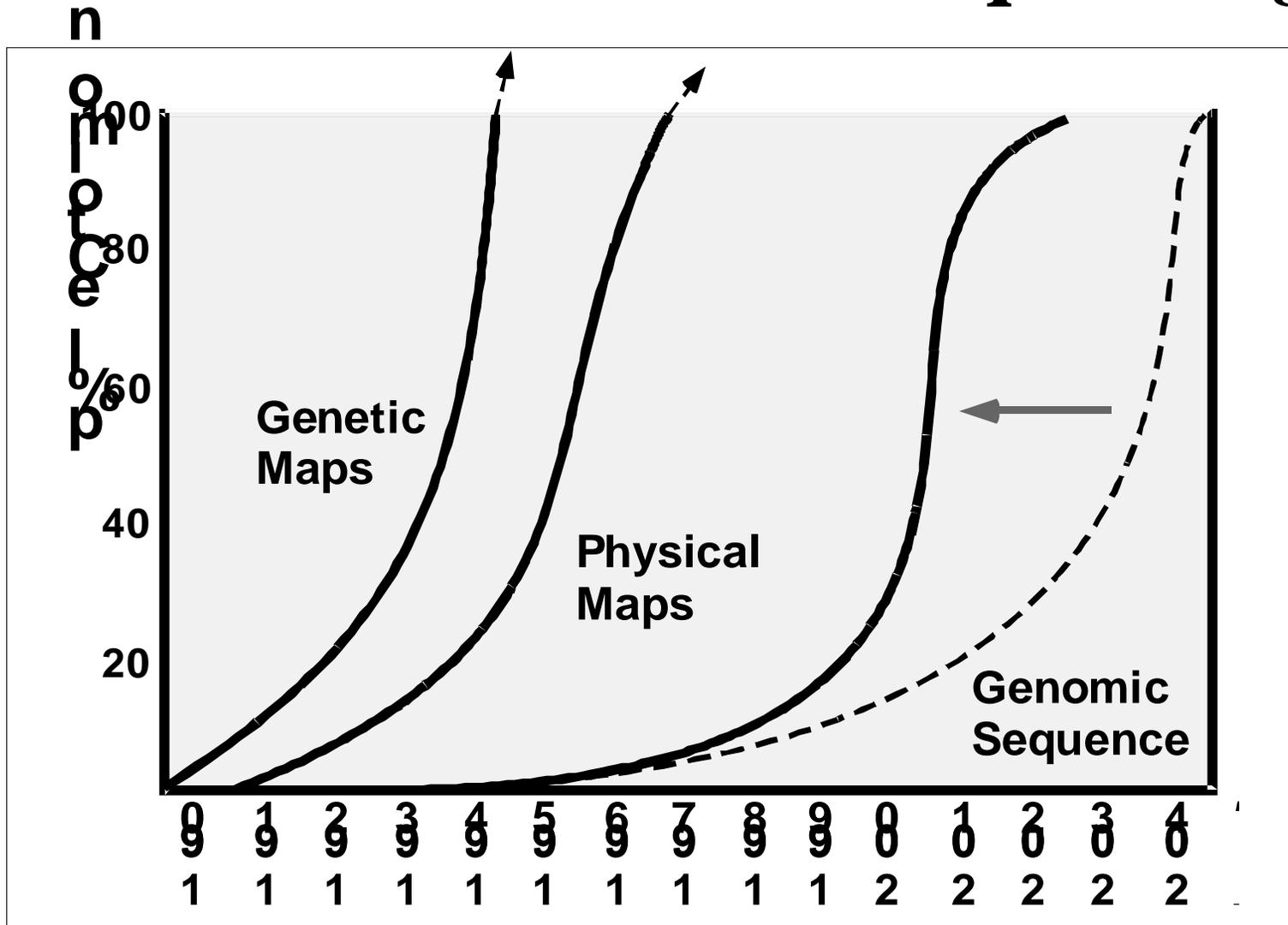
# **Human Genome Project: 5 Year Goals**

## **New Goals for the U.S. Human Genome Project: 1998–2003**

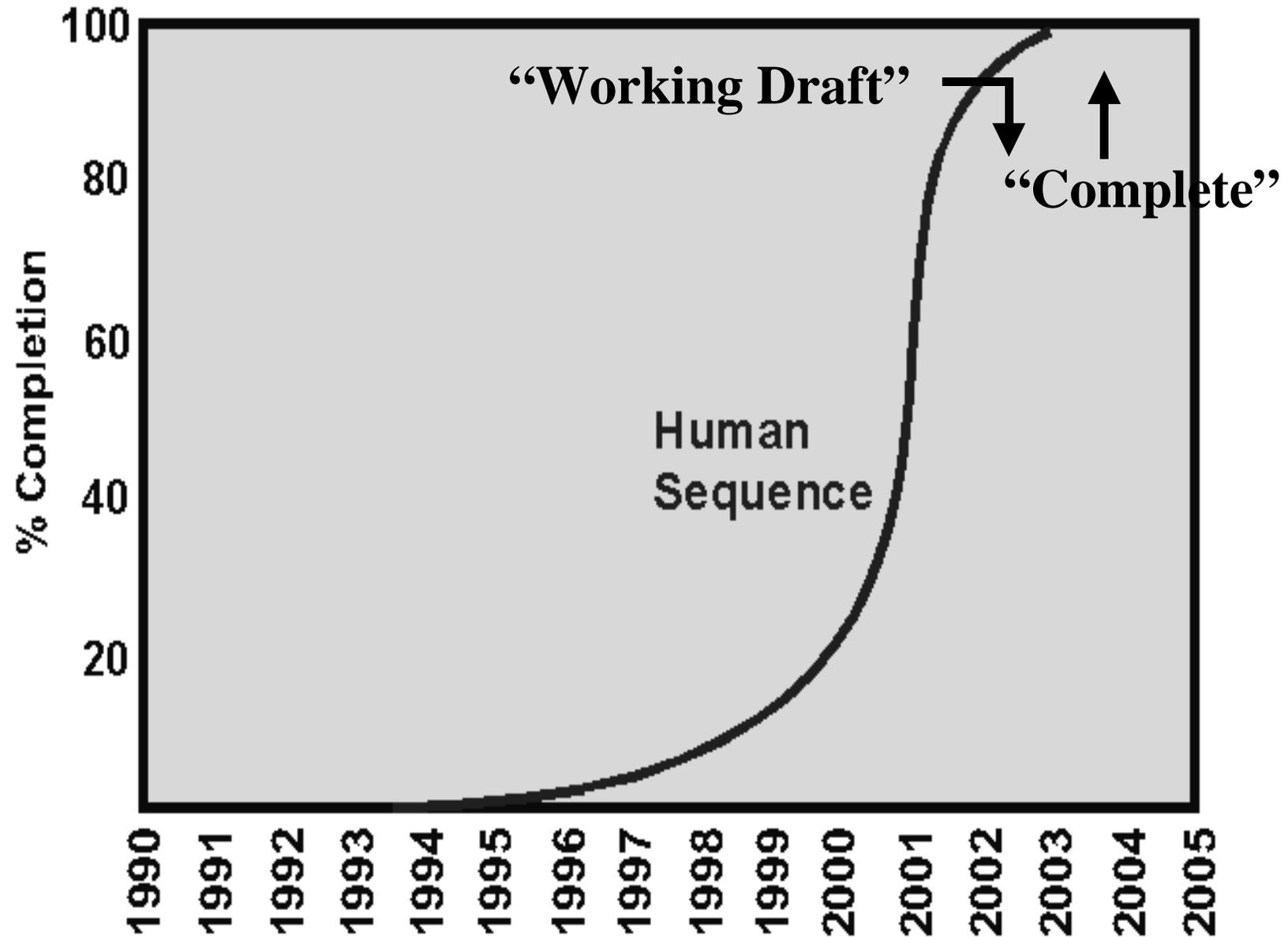
**Francis S. Collins,\* Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, and the members of the DOE and NIH planning groups**

***Science* 282:682-689, 1998**

# Revised Timetable for Sequencing



# Timetable for Human Genome Sequencing



# Quality/Utility of “Working Draft” Sequence

GENOME METHODS

## Analysis of the Quality and Utility of Random Shotgun Sequencing at Low Redundancies

John Bouck,<sup>1,3</sup> Webb Miller,<sup>2</sup> James H. Gorrell,<sup>1</sup> Donna Muzny,<sup>1</sup> and  
Richard A. Gibbs<sup>1</sup>

<sup>1</sup>Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas 77030 USA;

<sup>2</sup>Department of Computer Science and Engineering, Pennsylvania State University,  
University Park, Pennsylvania 16802 USA

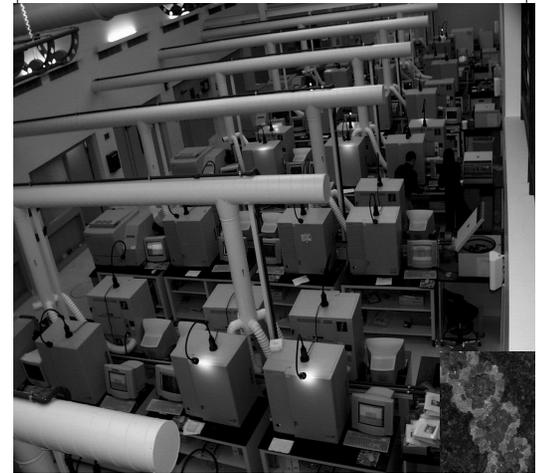
*Genome Research* 8:1074-1084, 1998

# Human Genome Sequencing Centers

**Genome Sequencing Center**  
Washington University  
School of Medicine  
St. Louis, MO USA



**Whitehead Institute/MIT  
Genome Sequencing Center**



**Baylor College of Medicine**



**The Sanger Centre**



**JGI**  
JOINT GENOME INSTITUTE



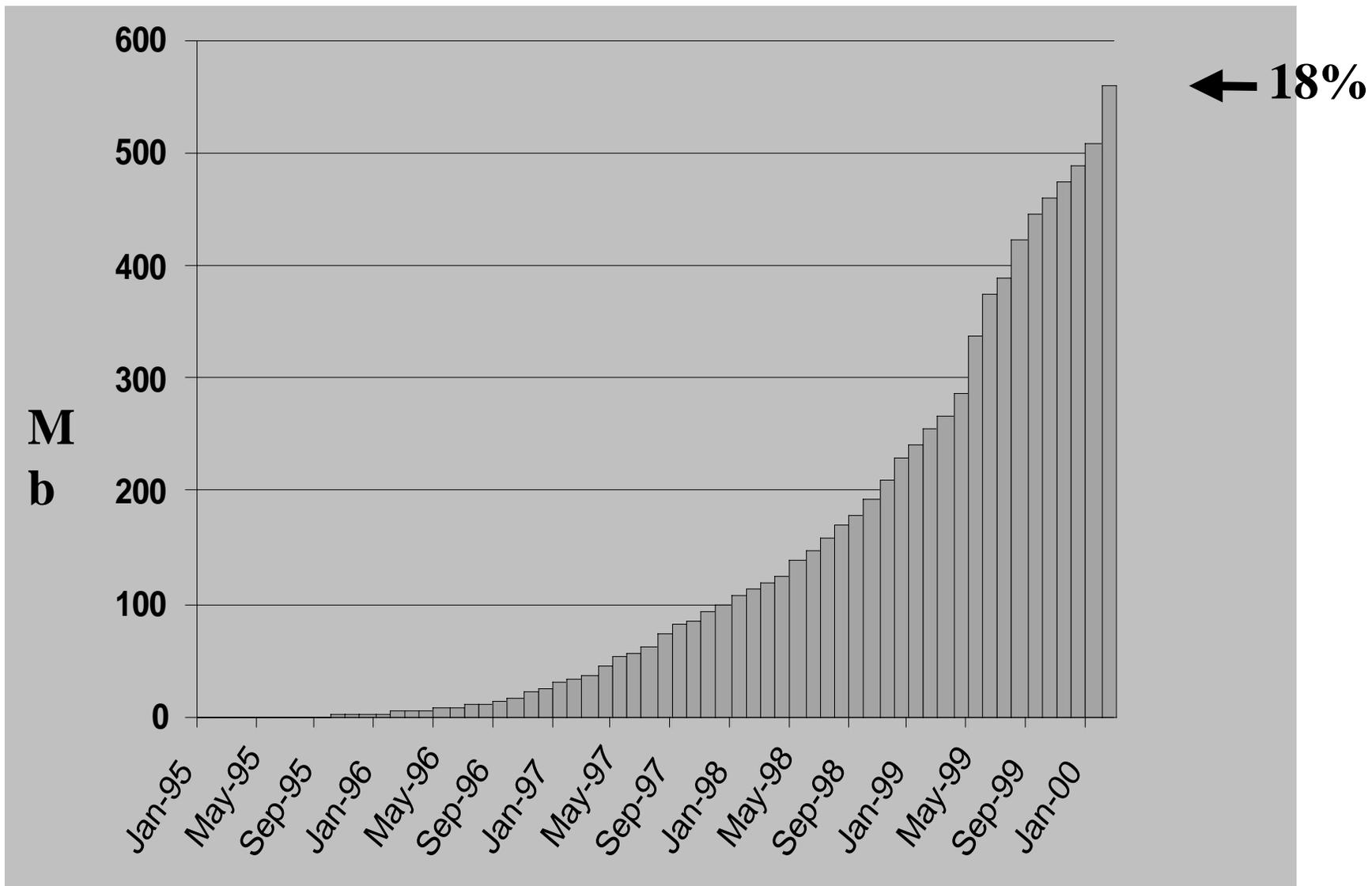
# Washington U. Genome Sequencing Center



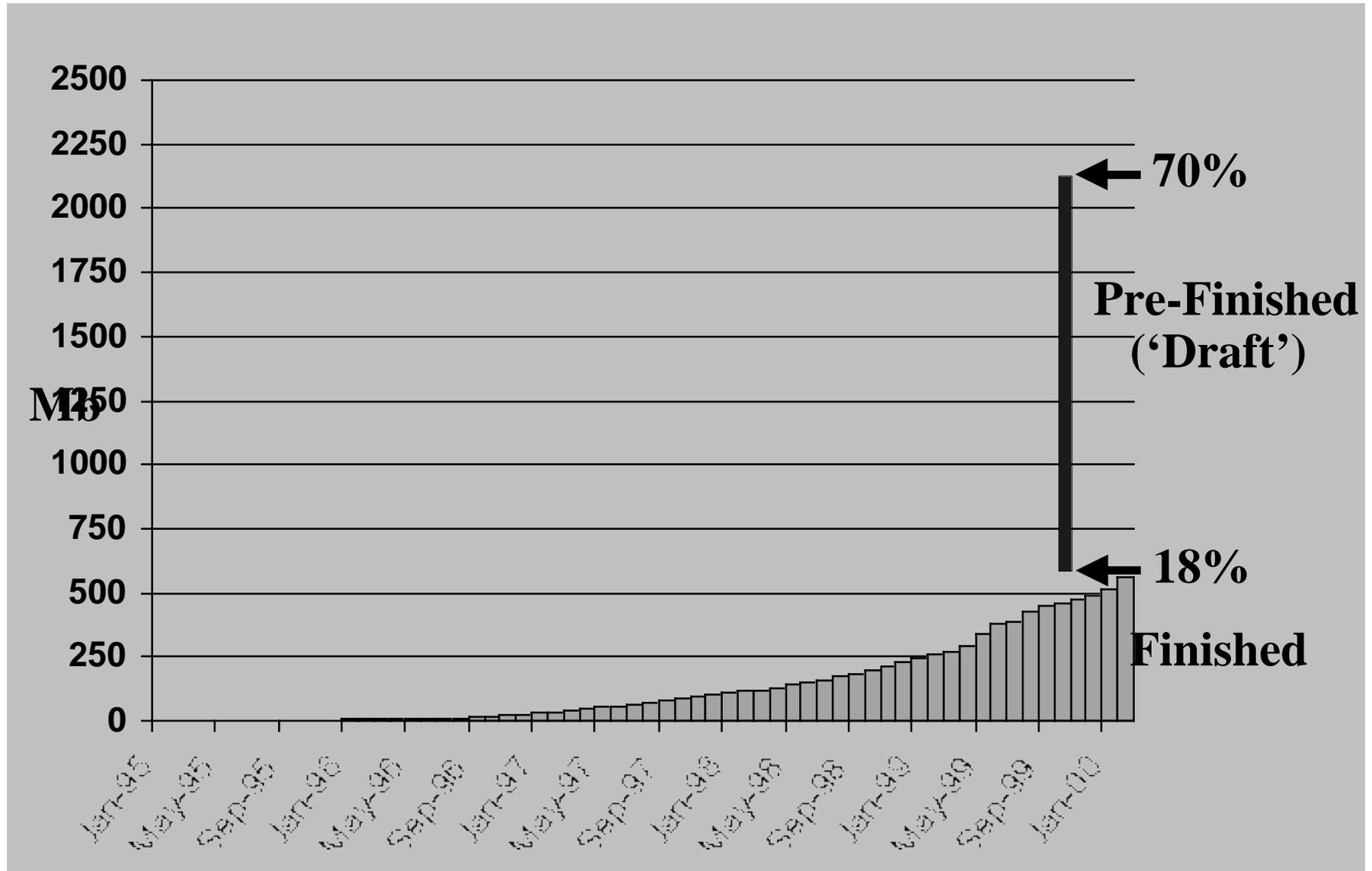
# Automation for Large-Scale Sequencing



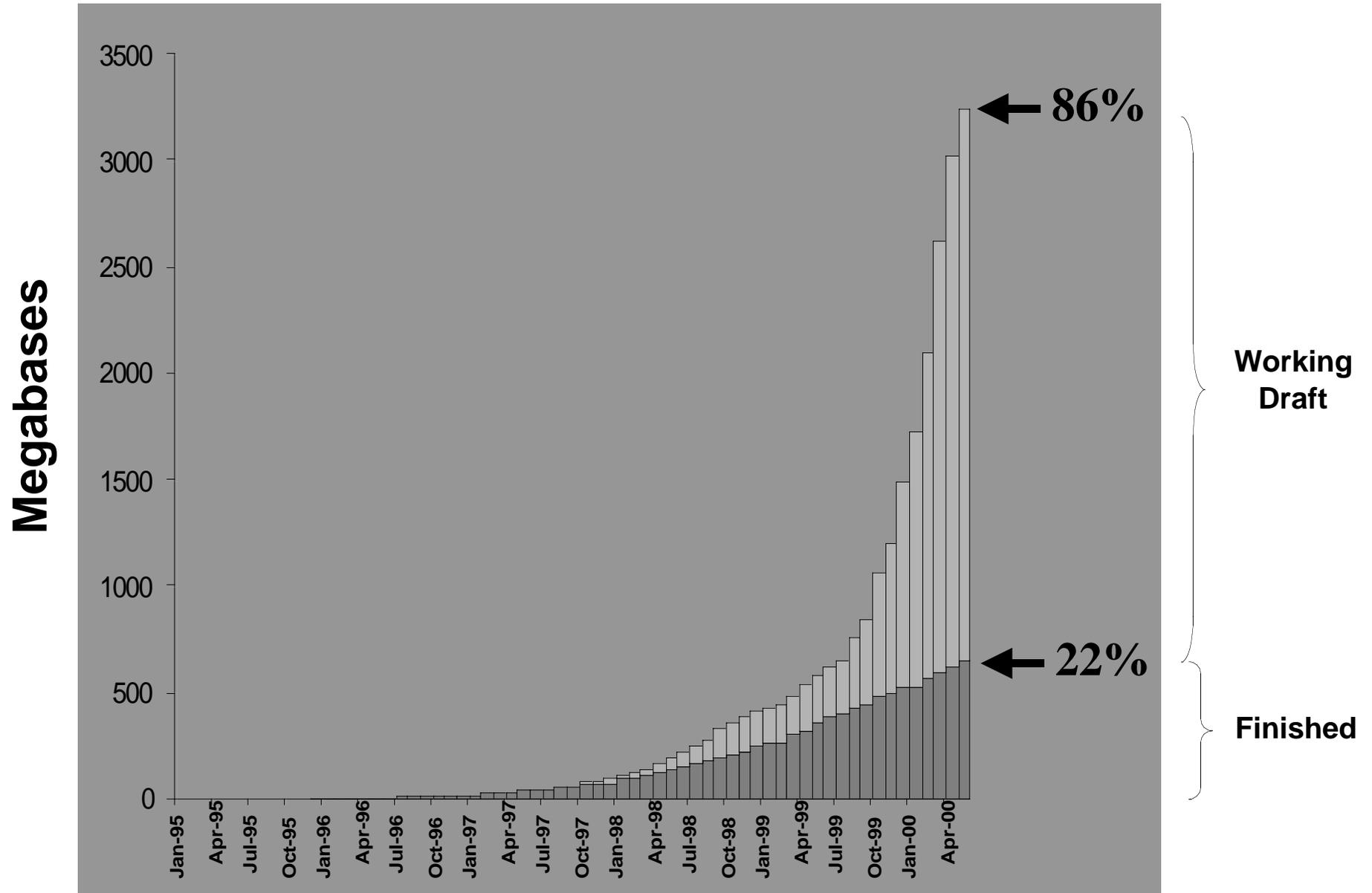
# Finished Human Genome Sequence



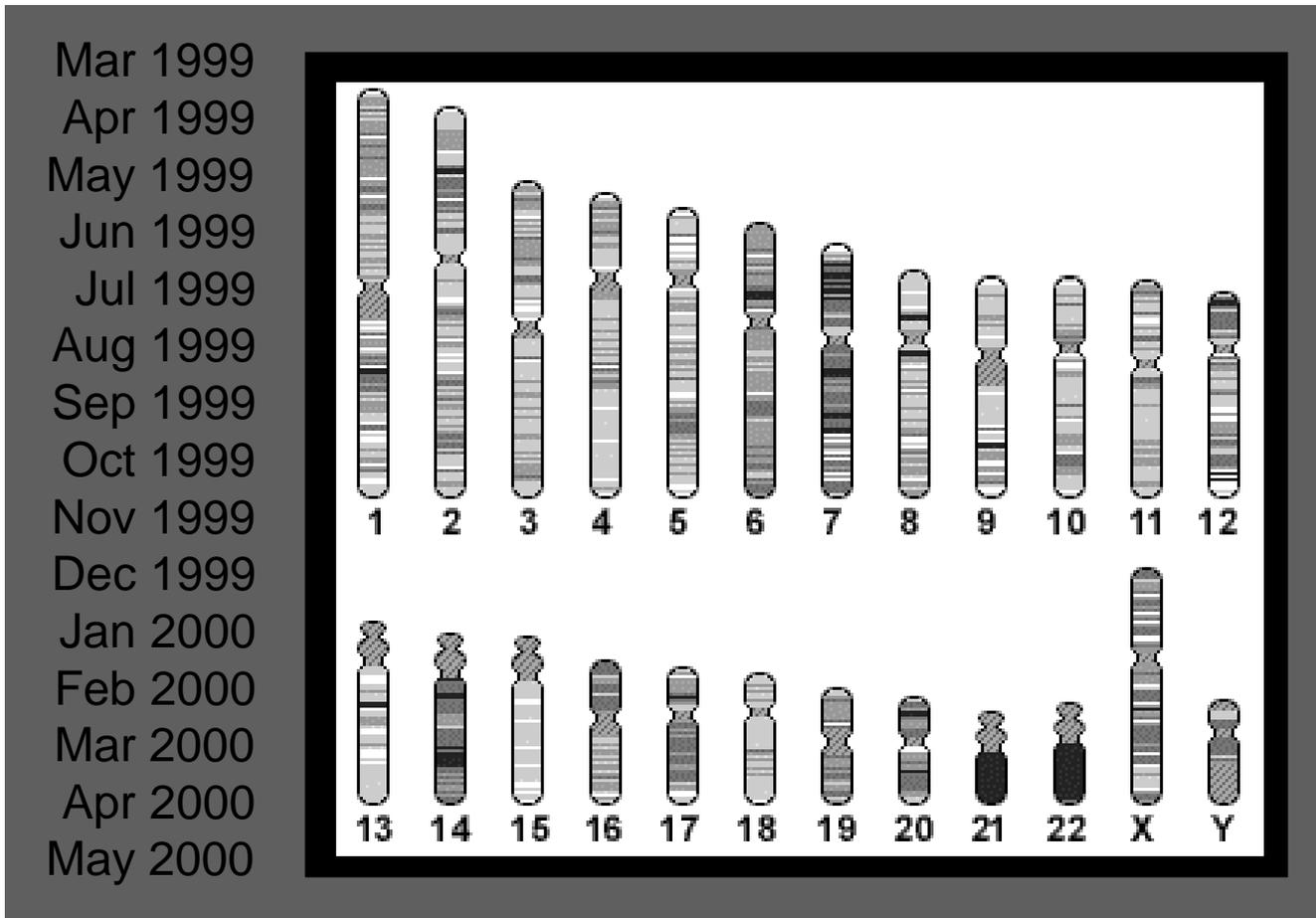
# Total Human Genome Sequence



# Human Genome Sequence



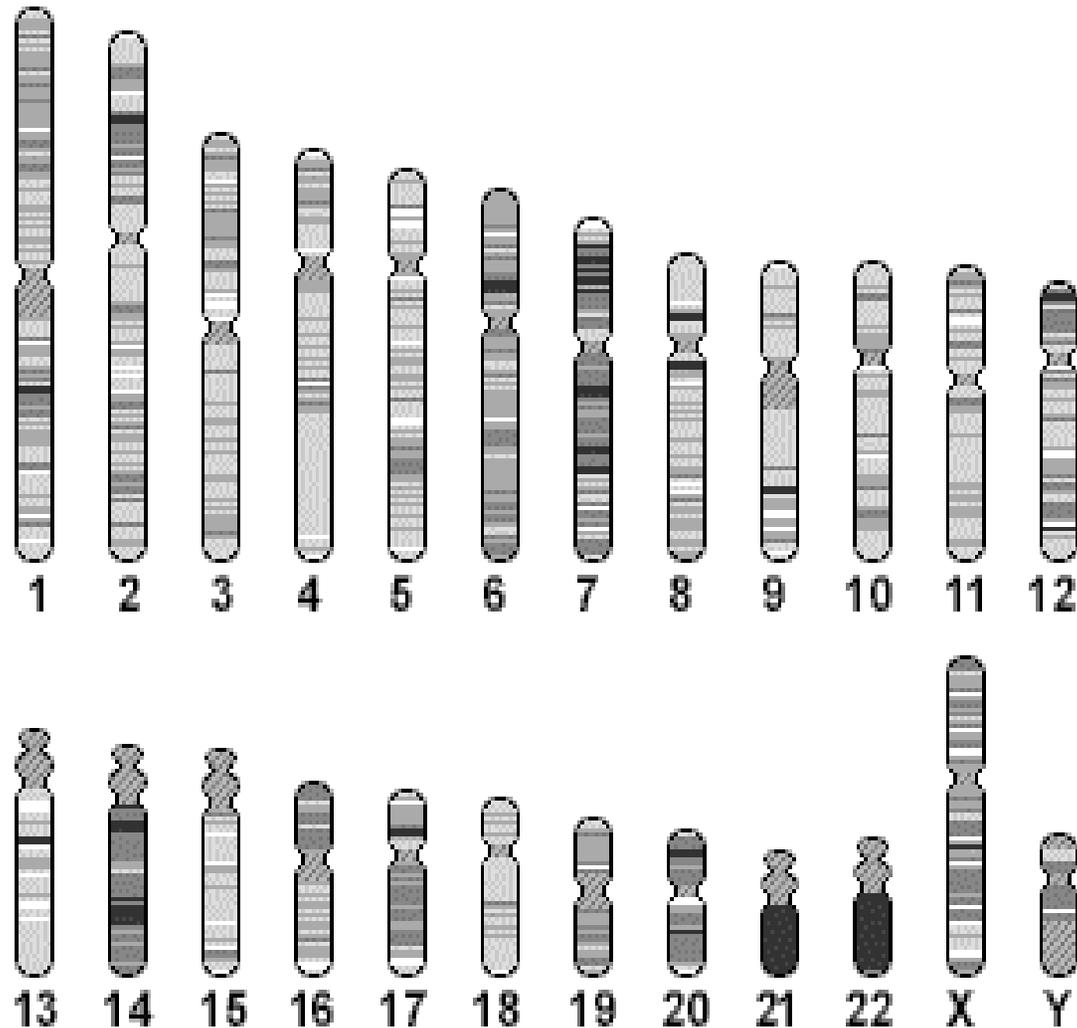
# Human Genome Sequencing



<http://www.ncbi.nlm.nih.gov/genome/seq>

# Human Sequencing: Current Progress

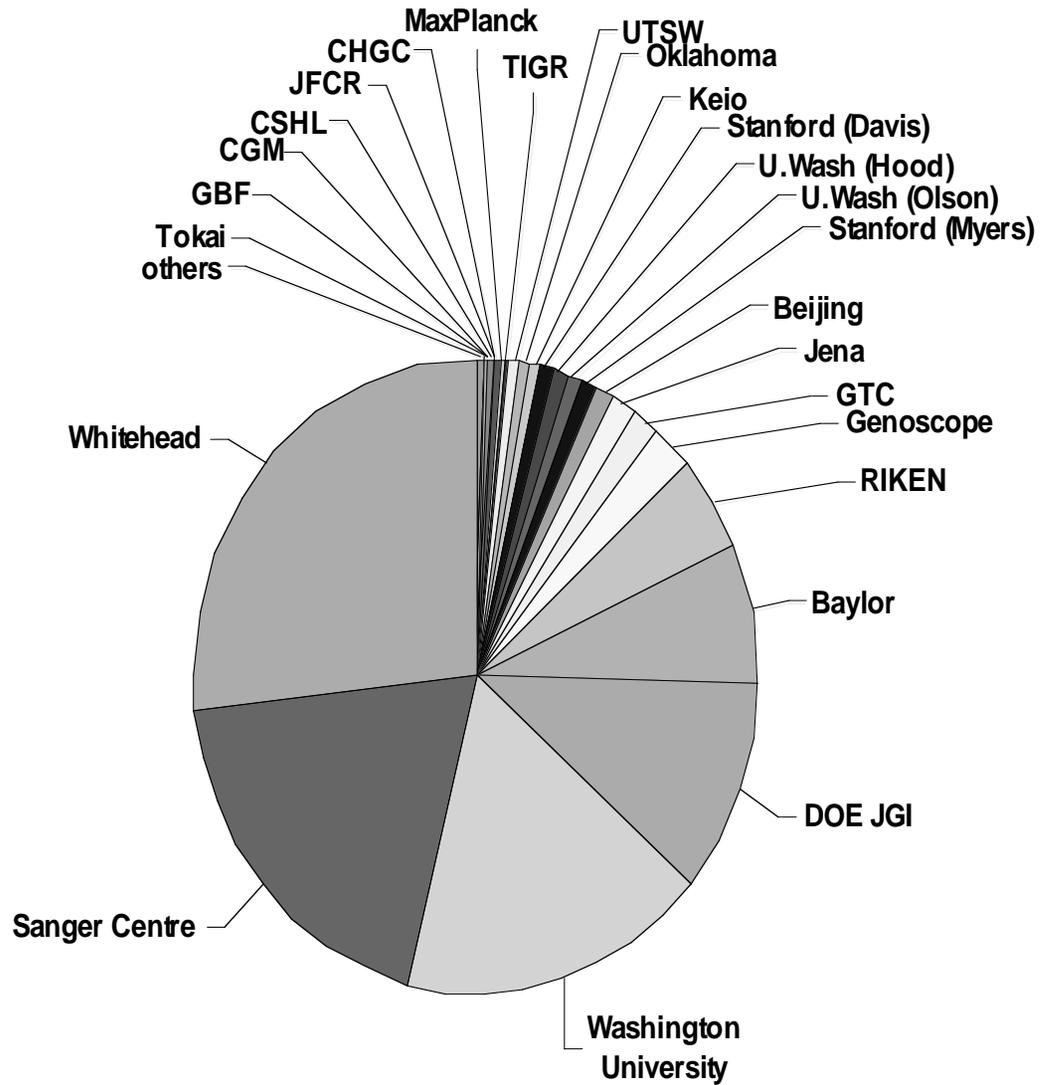
Sept. 18, 2000



**Draft: 66.2%**  
**Finished: 24.7%**  
**Total: 91.3%**

■ > 1000 kb   ■ 250-1000 kb   ■ < 250 kb  
□ draft sequence   ▨ heterochromatin

# Human Genome Sequencing



2 December 1999

International weekly journal of science

# nature

\$10.00

www.nature.com

## The first human chromosome sequence

.....  
**Climate change**  
Thermohaline trigger

.....  
**Intermolecular energetics**  
Good vibrations

.....  
**Impacts of foreseeable science**  
Supplement with this issue

**New on the market**  
Lasers

## The DNA sequence of human chromosome 22

I. Dunham, N. Shimizu, B. A. Roe, S. Chissoe *et al.*†

† A full list of authors appears at the end of this paper

Knowledge of the complete genomic DNA sequence of an organism allows a systematic approach to defining its genetic components. The genomic sequence provides access to the complete structures of all genes, including those without known function, their control elements, and, by inference, the proteins they encode, as well as all other biologically important sequences. Furthermore, the sequence is a rich and permanent source of information for the design of further biological studies of the organism and for the study of evolution through cross-species sequence comparison. The power of this approach has been amply demonstrated by the determination of the sequences of a number of microbial and model organisms. The next step is to obtain the complete sequence of the entire human genome. Here we report the sequence of the euchromatic part of human chromosome 22. The sequence obtained consists of 12 contiguous segments spanning 33.4 megabases, contains at least 545 genes and 134 pseudogenes, and provides the first view of the complex chromosomal landscapes that will be found in the rest of the genome.

*Nature* 402:489-495, 1999

18 May 2000

International weekly journal of science

# nature

510.00

www.nature.com



## Counting down from 21

**Optical microscopy** A molecular light-bulb

**Bioluminescence** Jellyfish blues

**Proglacial lakes** Breaking their own dam

**nature jobs focus**

Postgraduate opportunities

## The DNA sequence of human chromosome 21

The chromosome 21 mapping and sequencing consortium

M. Hattori<sup>1</sup>\*, A. Fujiyama<sup>2</sup>, T. D. Taylor<sup>3</sup>, H. Watanabe<sup>4</sup>, T. Yada<sup>5</sup>, H.-S. Park<sup>6</sup>, A. Toyoda<sup>7</sup>, K. Ishii<sup>8</sup>, Y. Totoki<sup>9</sup>, D.-K. Choi<sup>10</sup>, E. Soeda<sup>11</sup>, M. Ohki<sup>12</sup>, T. Takagi<sup>13</sup>, Y. Sakaki<sup>14</sup>; S. Taudien<sup>15</sup>\*, K. Blechschmidt<sup>16</sup>, A. Polley<sup>17</sup>, U. Menzel<sup>18</sup>, J. Delabar<sup>19</sup>, K. Kumpf<sup>20</sup>, R. Lehmann<sup>21</sup>, D. Patterson<sup>22</sup>, K. Reichwald<sup>23</sup>, A. Rump<sup>24</sup>, M. Schilhabell<sup>25</sup>, A. Schudy<sup>26</sup>, W. Zimmermann<sup>27</sup>, A. Rosenthal<sup>28</sup>; J. Kudoh<sup>29</sup>\*, K. Shibuya<sup>30</sup>, K. Kawasaki<sup>31</sup>, S. Asakawa<sup>32</sup>, A. Shintani<sup>33</sup>, T. Sasaki<sup>34</sup>, K. Nagamine<sup>35</sup>, S. Mitsuyama<sup>36</sup>, S. E. Antonarakis<sup>37</sup>\*, S. Minooshima<sup>38</sup>, N. Shimizu<sup>39</sup>; G. Nordtsiek<sup>40</sup>\*, K. Hornischer<sup>41</sup>, P. Brandt<sup>42</sup>, M. Scharfe<sup>43</sup>, O. Schön<sup>44</sup>, A. Desario<sup>45</sup>, J. Reichelt<sup>46</sup>, G. Kauer<sup>47</sup>, H. Blöcker<sup>48</sup>; J. Ramser<sup>49</sup>\*, A. Beck<sup>50</sup>, S. Klages<sup>51</sup>, S. Hennig<sup>52</sup>, L. Rieselmann<sup>53</sup>, E. Dagand<sup>54</sup>, S. Wehrmeyer<sup>55</sup>, K. Borzyn<sup>56</sup>, K. Gardiner<sup>57</sup>, D. Nizel<sup>58</sup>, F. Francis<sup>59</sup>, H. Lehrach<sup>60</sup>, R. Reinhardt<sup>61</sup> & M.-L. Yaspo<sup>62</sup>

*Nature* 405:311-319, 2000

# Human Genome Sequence by the HGP

- **Immediate Release**

Sequence Contigs >1-2 kb

Finished and Pre-Finished Sequence

- **High Accuracy**

Error Rate of <1 in 10,000 bp

Assessed/Confirmed by QC Exercises  
(see *Genome Research* 9:1-4, 1999)

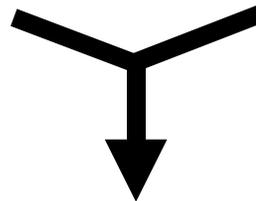
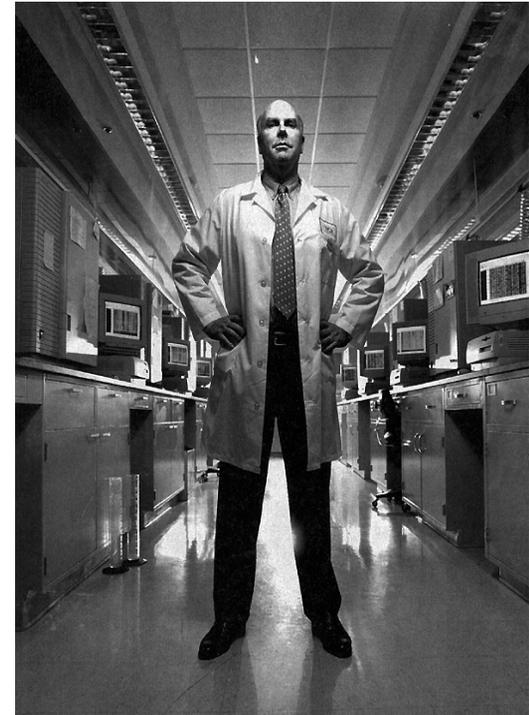
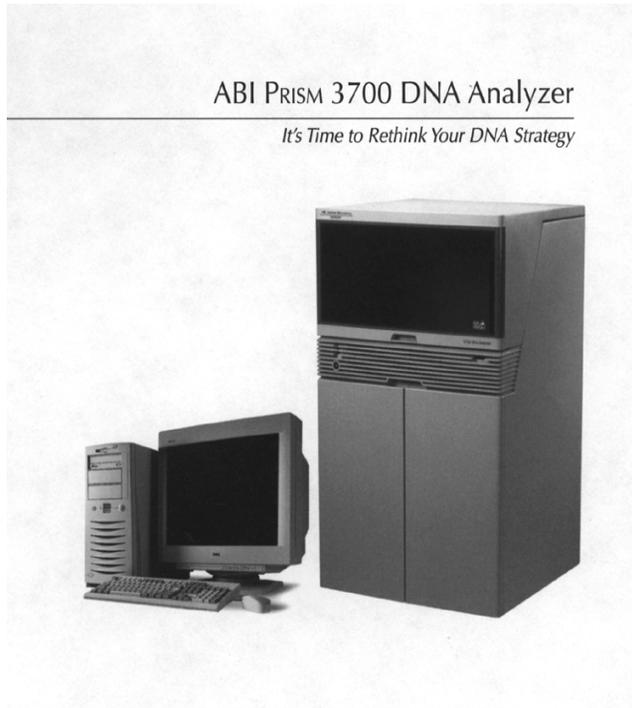
- **Cost**

Steady (But Not Massive) Decrease

Currently at ~25-50¢ per Finished bp

# **The Private Sector and DNA Sequencing**

# Commercial Interest in Human Genome Sequencing



**CELERA**



# Whole-genome Shotgun Sequencing

## **Pros:**

(Weber and Myers, 1997)

**No sequence-ready maps required!**  
Savings of effort and cost

**Much faster than clone-by-clone**

**Detection of DNA polymorphisms**  
Sequence 5 individuals

## **Cons:**

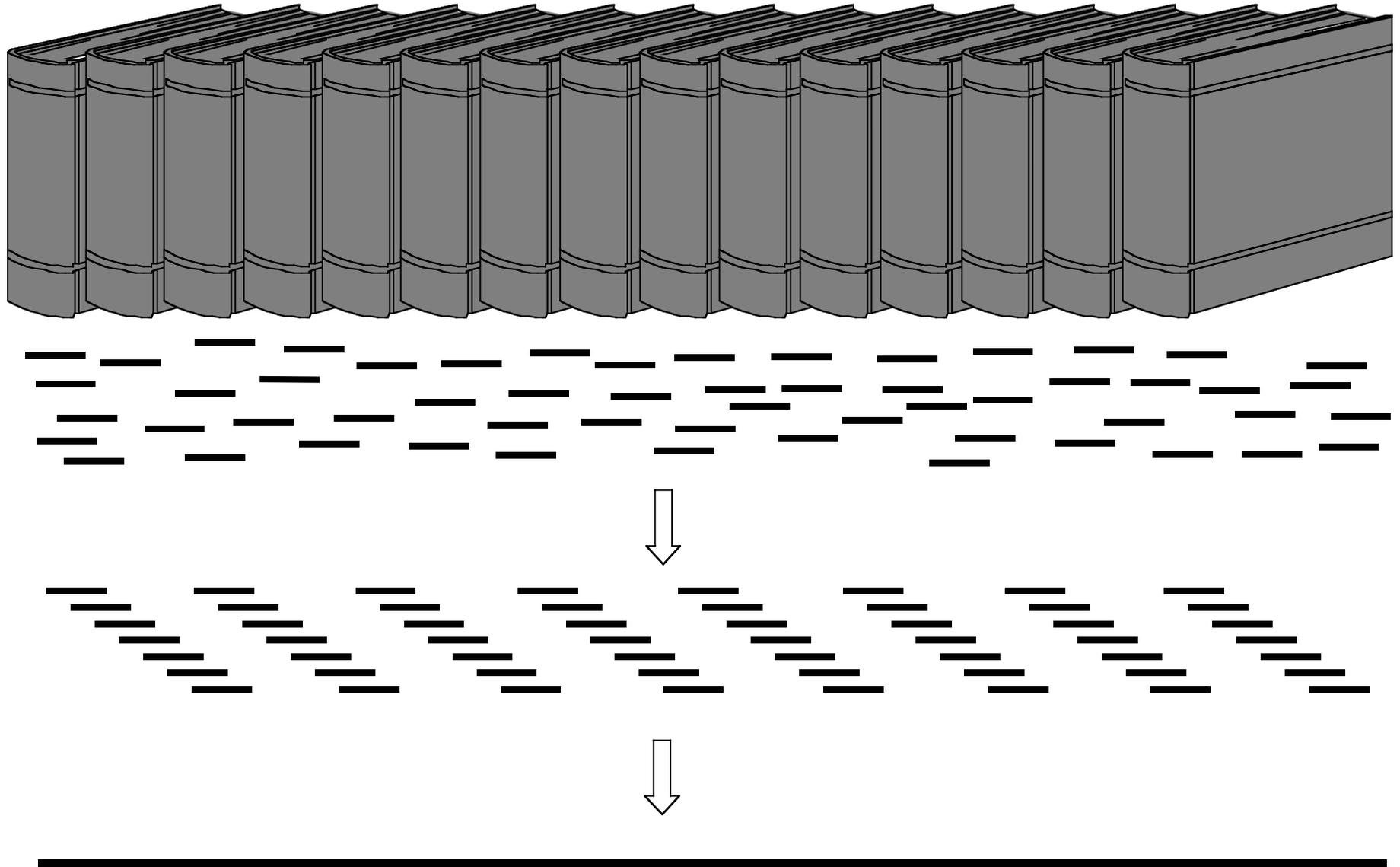
(Green, 1997)

**Probability of success debatable**

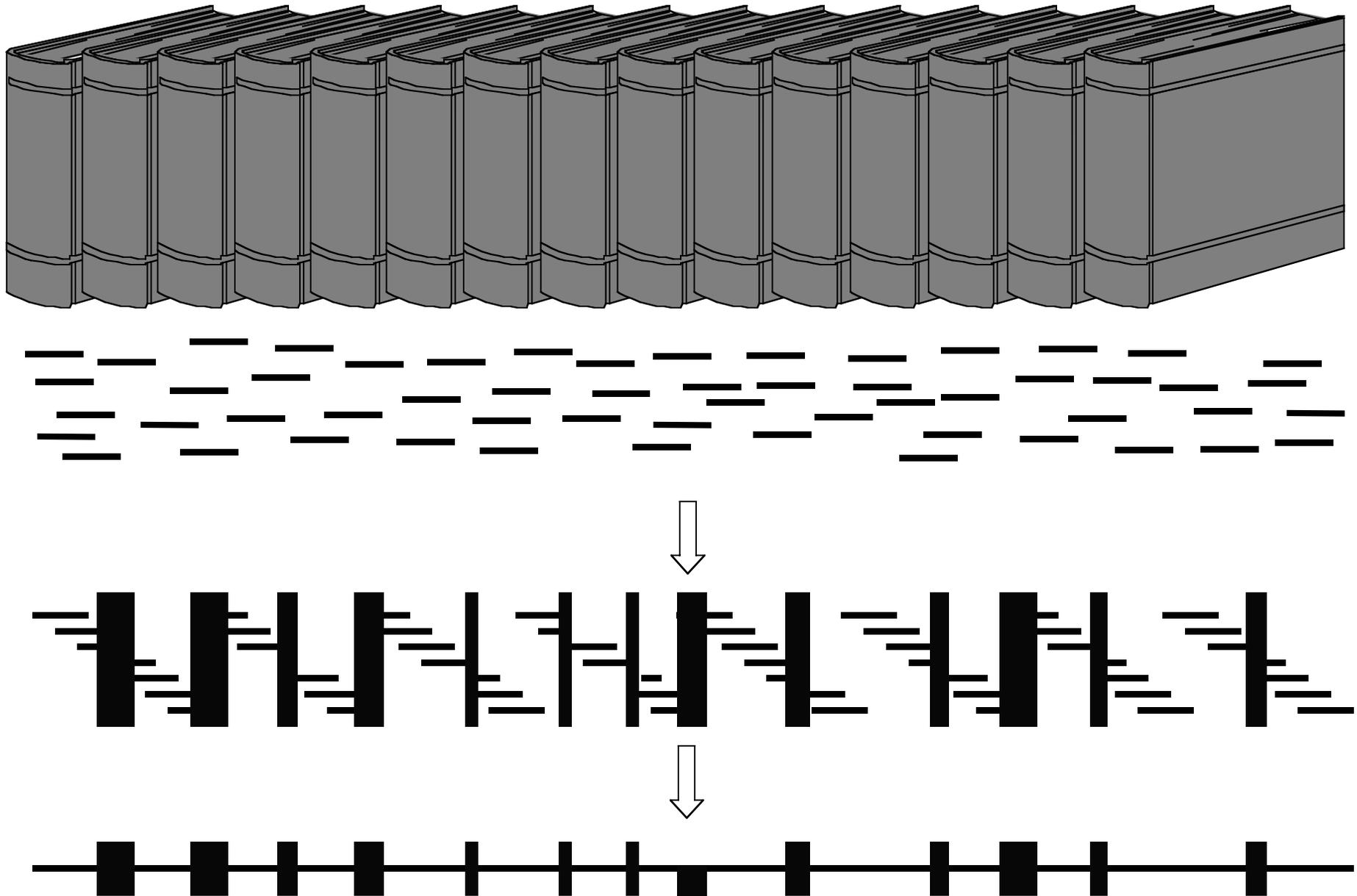
**Cost savings ?**  
(“Finishing” could be a mess)

**Final quality level ?**

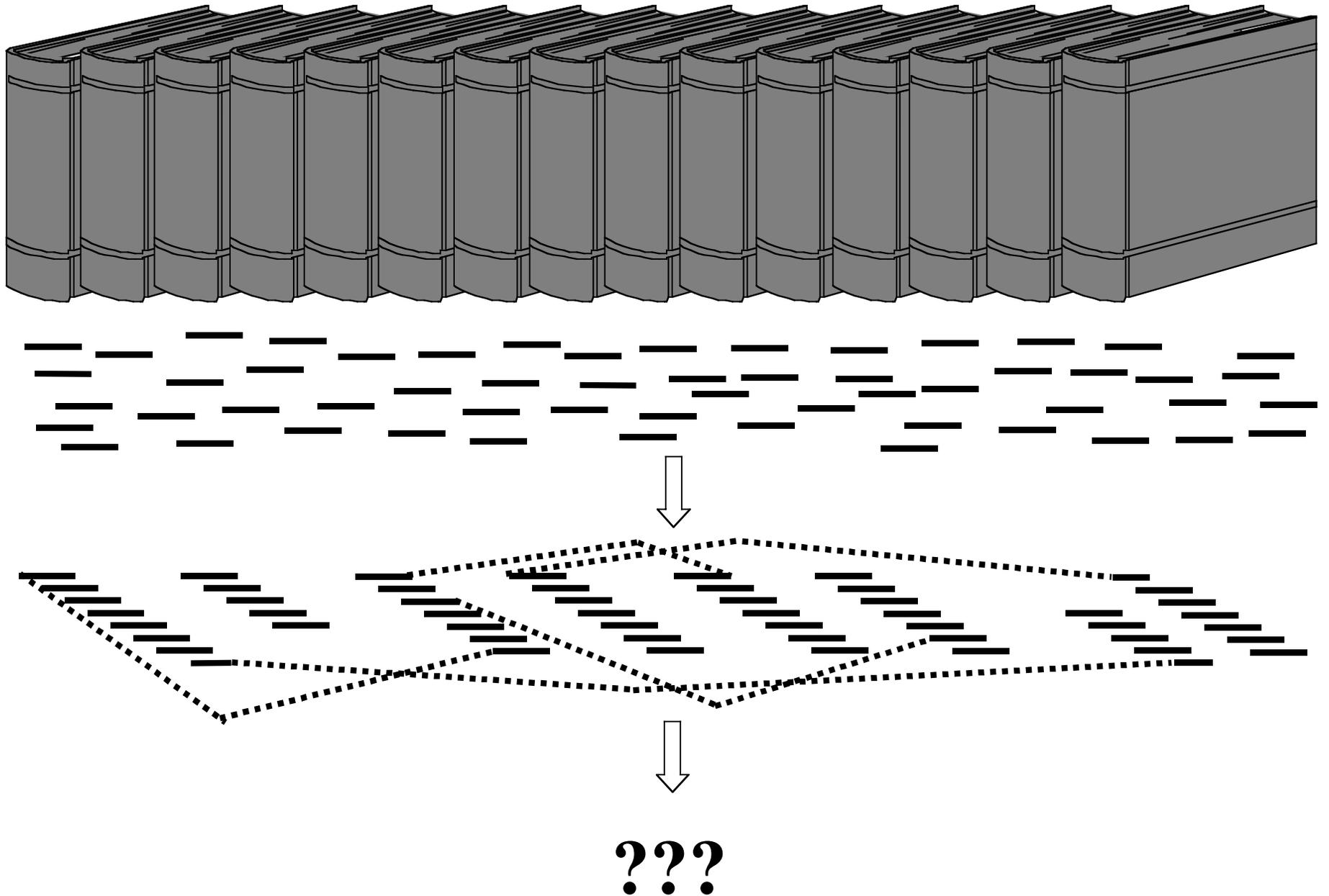
# Whole Genome Shotgun Sequencing



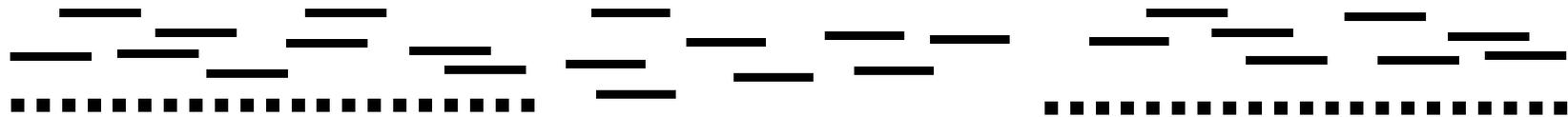
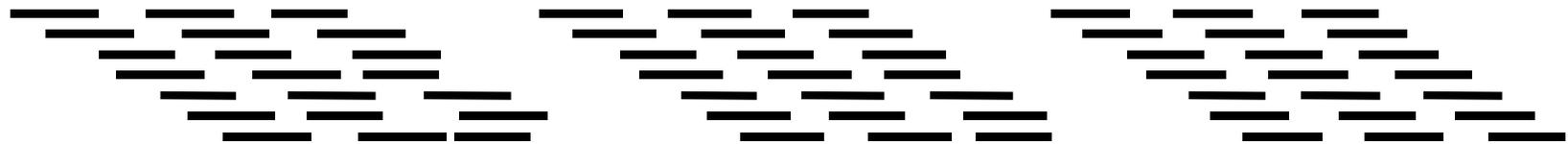
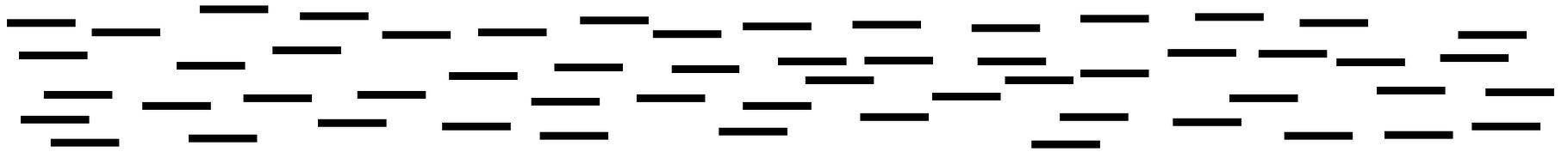
# Whole Genome Shotgun Sequencing



# Whole Genome Shotgun Sequencing



# Hybrid Sequencing Strategy



**BAC 1**

**BAC 2**

**BAC 3**

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

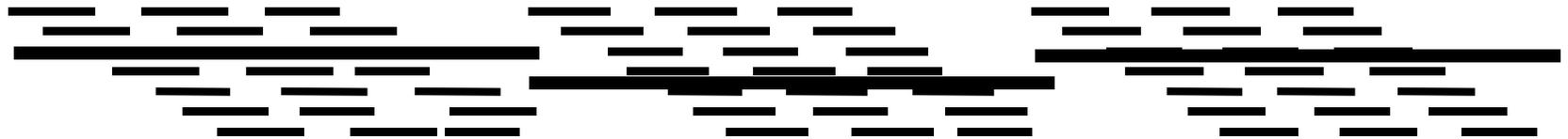
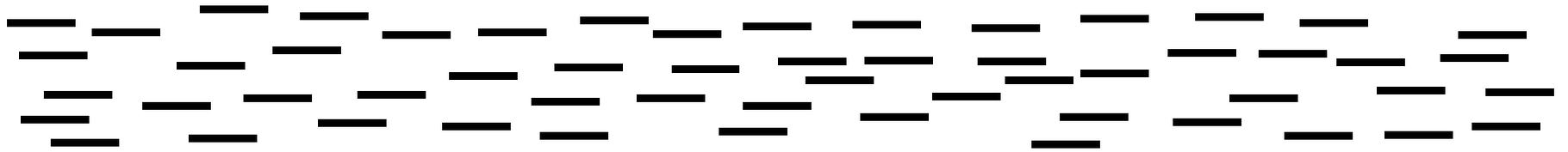
GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

# Hybrid Sequencing Strategy



**BAC 1**

**BAC 2**

**BAC 3**

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

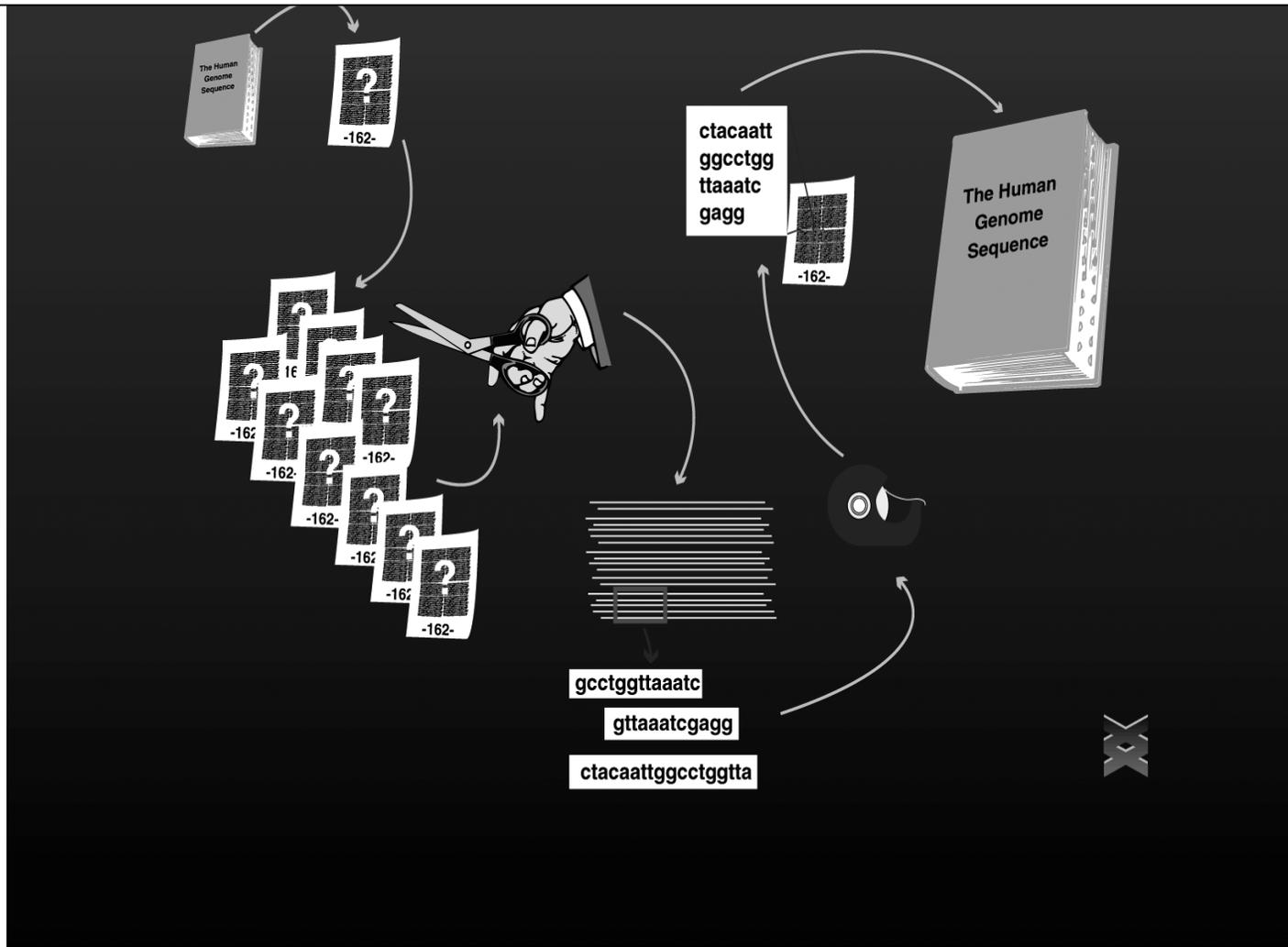
GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

```
GATCGTCTAGAATCTC
GAGATCTCTGAGAGTC
GTGGGAAACTGTGTGA
TGTGACTAGCCACAGT

TACGTGTGAGAGATGT
ATGATGCACCTGACCC
GGGTTTCACTCTCAAC
GACTCACTCCACCTCA

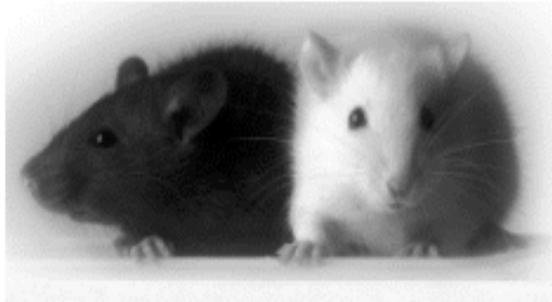
GAGGCCACCGCCGCT
GTGCACGTCCACCACC
```

# Sequencing Mapped DNA



# **Sequencing Other Genomes**

# Mouse Genome Analysis



## An action plan for mouse genomics

James Battey<sup>1</sup>, Elke Jordan<sup>2</sup>, David Cox<sup>3</sup> & William Dove<sup>4</sup>

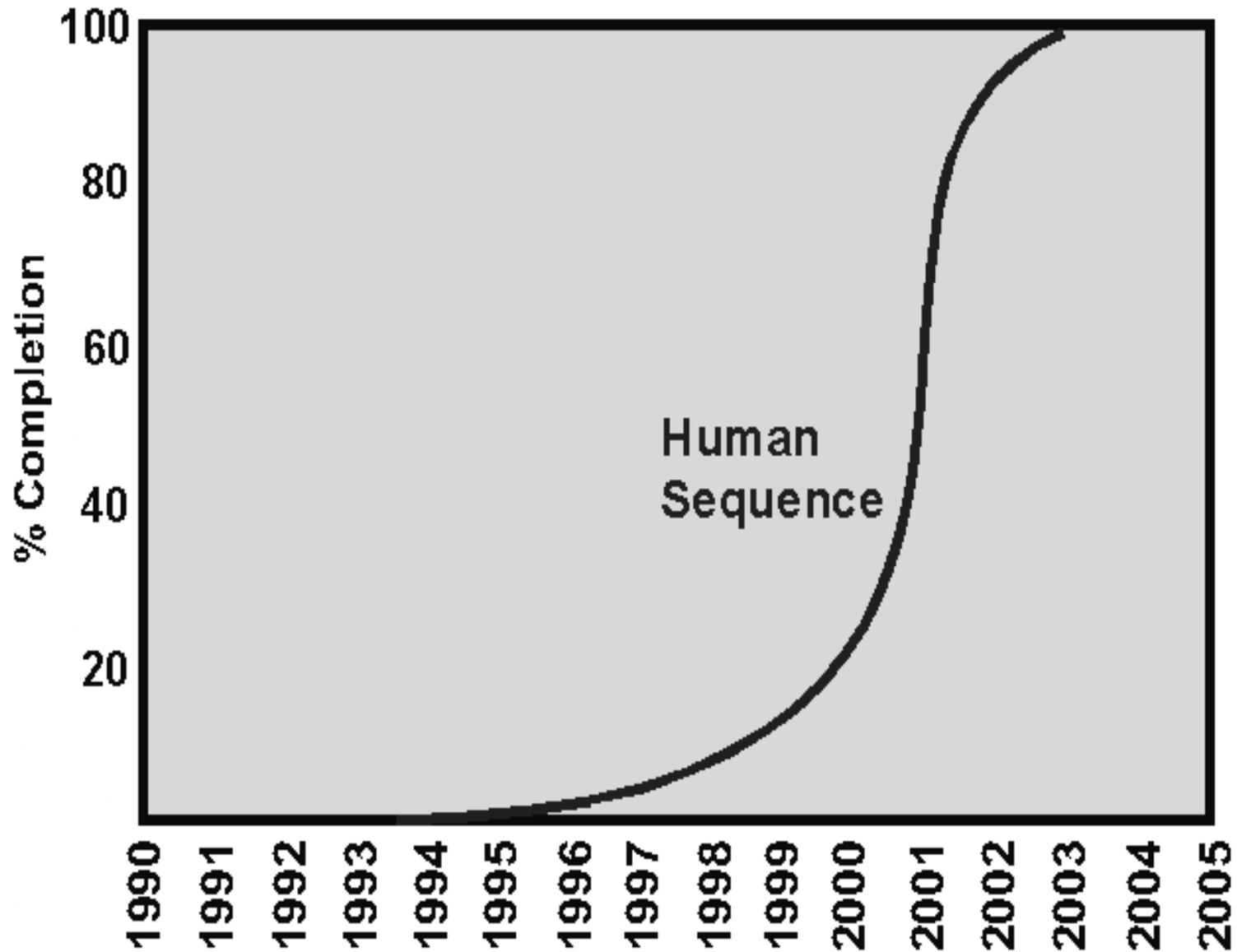
---

The mouse has become the leading animal model for studying biological processes in mammals. Creation of additional genomic and genetic resources will make the mouse an even more useful model for the research community. On the basis of recommendations from the scientific community, the National Institutes of Health (NIH) plans to support grants to generate a 'working draft' sequence of the mouse genome by 2003, systematic mutagenesis and phenotyping centres, repositories for mouse strain maintenance, distribution and cryopreservation and training fellowships in mouse pathobiology.

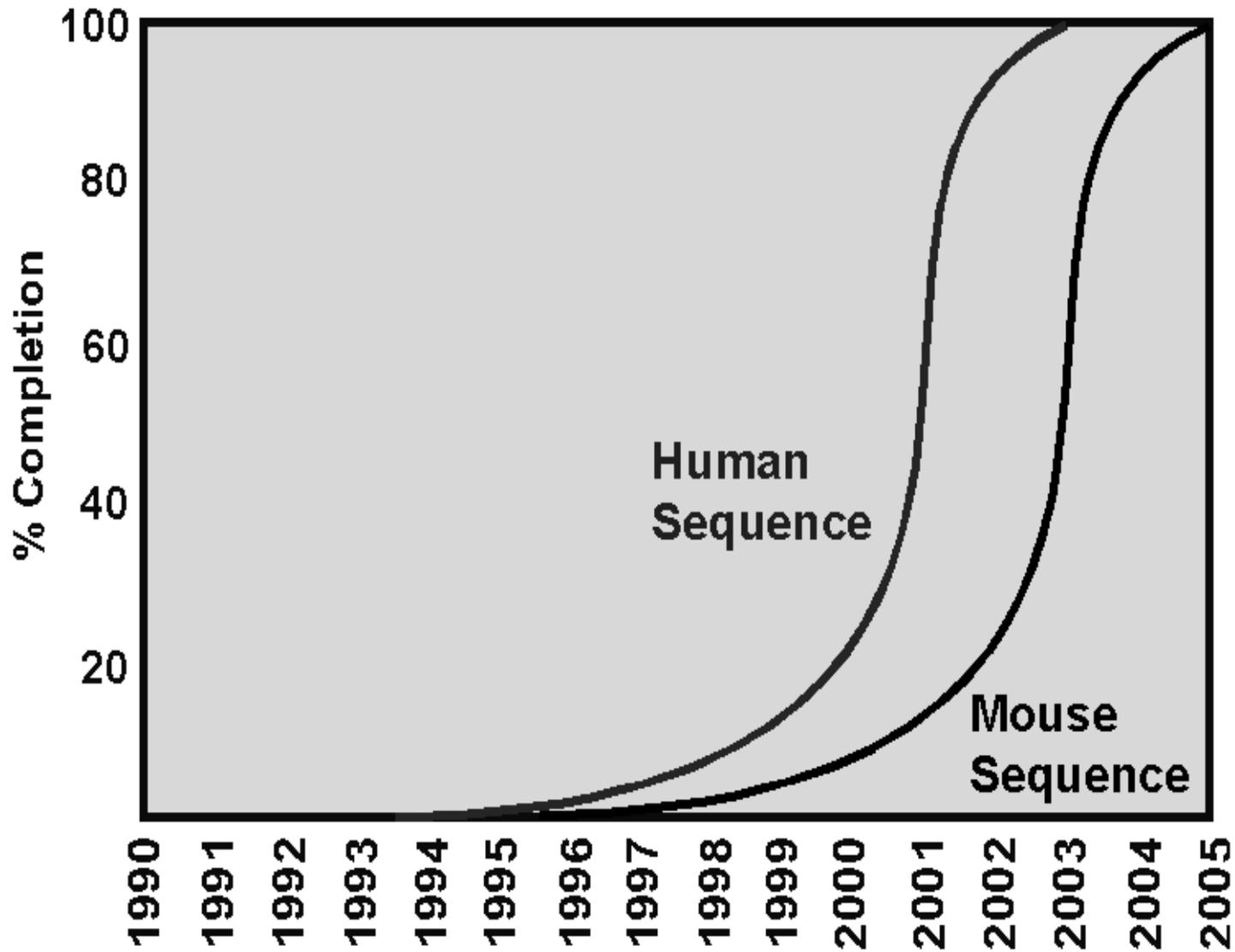
---

*Nature Genetics* 21:73-75, 1999  
<http://www.nih.gov/science/mouse>

# Highly Ambitious...



# Even More Ambitious...



# **NIH Plan for Mouse Genome Sequencing**

- **Consortium of Sequencing Centers**

  - Mapping Component: Fingerprints, End Sequences**

  - Sequencing Component: Global and Targeted Efforts**

- **Global Sequencing Efforts: “Hybrid Strategy”**

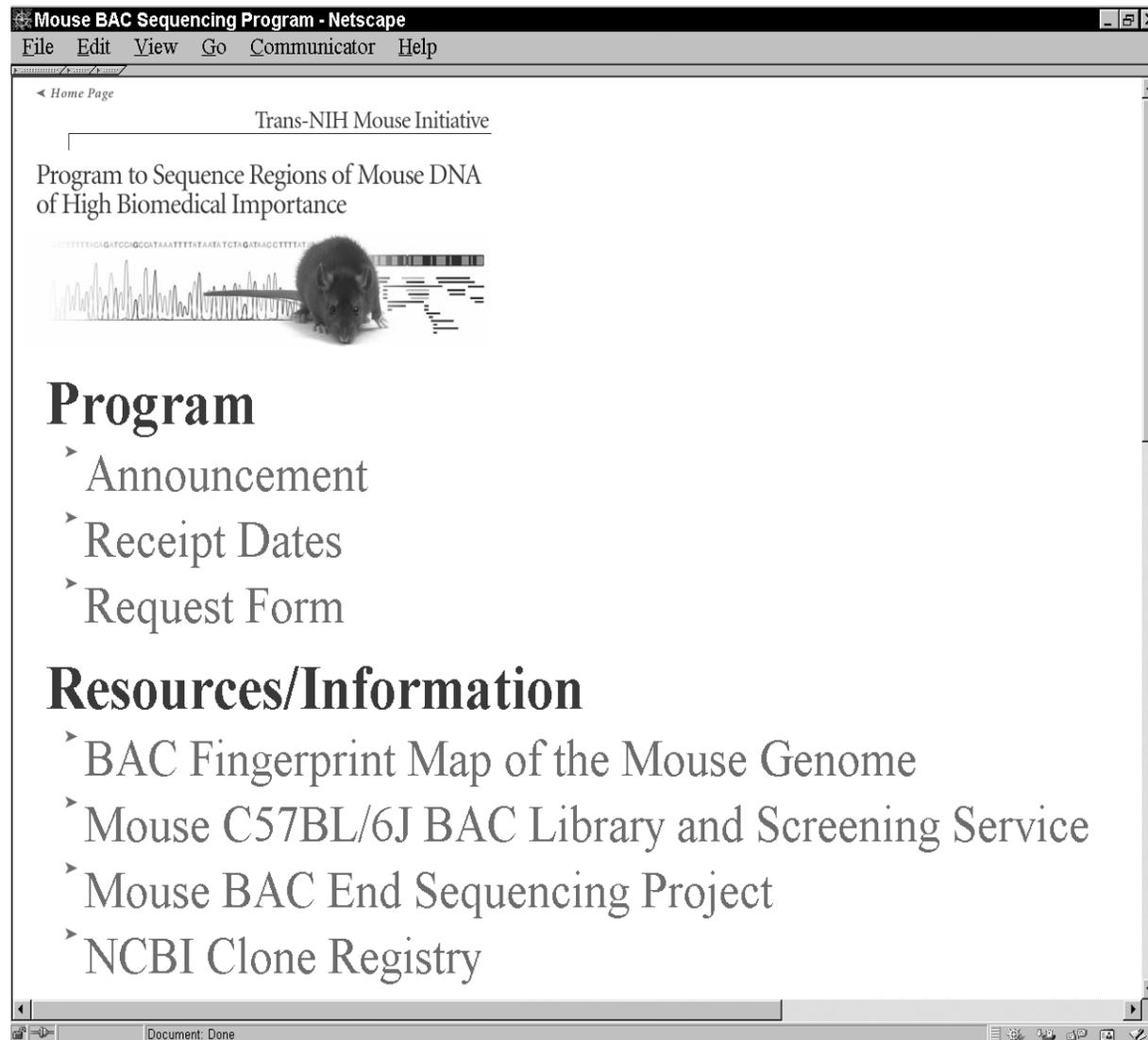
  - High Redundancy Whole-Genome-Shotgun Sequencing**

  - Low Redundancy BAC-by-BAC Sequencing**

- **Targeted Sequencing Efforts**

  - Prioritized Sequencing of Regions of Biomedical Importance**

# Prioritized Sequencing of Regions of Biomedical Importance



Mouse BAC Sequencing Program - Netscape

File Edit View Go Communicator Help

< Home Page

Trans-NIH Mouse Initiative

Program to Sequence Regions of Mouse DNA  
of High Biomedical Importance

TTTTTTHGATCGGCGTAAATTTTAAATCTAGATACCTTTTAT

**Program**

- Announcement
- Receipt Dates
- Request Form

**Resources/Information**

- BAC Fingerprint Map of the Mouse Genome
- Mouse C57BL/6J BAC Library and Screening Service
- Mouse BAC End Sequencing Project
- NCBI Clone Registry

Document: Done

<http://www.nih.gov/science/models/mouse/mouseseq>

# Annotating the Human “Working Draft” with Mouse Sequence

*progress*

---

## **Shotgun sample sequence comparisons between mouse and human genomes**

John B. Bouck, Michael L. Metzker & Richard A. Gibbs

*Nature Genetics* 25:31-33, 2000

# Importance of Comparative Sequence Analysis

